PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA

Universidade Federal do Rio de Janeiro

Diogo Nolasco Ferreira Sousa

# IDENTIFICAÇÃO AUTOMÁTICA DE ÁREAS DE PESQUISA EM C&T

**DISSERTAÇÃO DE MESTRADO**

Rio de Janeiro

Instituto de Matemática

Instituto Tércio Pacitti de Aplicações e Pesquisas Computacionais

UNIVERSIDADE FEDERAL DO RIO DE JANEIRO
INSTITUTO DE MATEMÁTICA
INSTITUTO TÉRCIO PACITTI DE APLICAÇÕES E PESQUISAS COMPUTACIONAIS
PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA

Diogo Nolasco Ferreira Sousa

# IDENTIFICAÇÃO AUTOMÁTICA DE ÁREAS DE PESQUISA EM C&T

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Informática, Instituto de Matemática e Instituto Tércio Pacciti, Universidade Federal do Rio de Janeiro, como requisito parcial à obtenção do título de Mestre em Informática.

Orientador: Prof. Jonice de Oliveira Sampaio, D.Sc.

Rio de Janeiro
2016

Diogo Nolasco Ferreira Sousa

# IDENTIFICAÇÃO AUTOMÁTICA DE ÁREAS DE PESQUISA EM C&T

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Informática, Instituto de Matemática e Instituto Tércio Pacciti, Universidade Federal do Rio de Janeiro, como requisito parcial à obtenção do título de Mestre em Informática.

Aprovada em 30/11/2016.

_____

Prof. Jonice de Oliveira Sampaio, D.Sc., UFRJ

_____

Prof. Maria Luiza Machado Campos, Ph.D., UFRJ

_____

Prof. Adelaide Maria de Souza Antunes, D.Sc., INPI

_____

Marcia de Freitas Lenzi, D.Sc., Fiocruz

# Agradecimentos

Agradeço primeiramente a Deus por ter me dado forças e saúde nas dificuldades e momentos em que eu mais precisei. Agradeço aos meus pais, Luciara e Ginaldo e a minha irmã Luciana, pelo apoio emocional, pelo grande incentivo e por terem me dado condições de chegar até aqui. Agradeço também ao PPGI e a UFRJ por terem me dado o conhecimento e as ferramentas necessárias para realizar esse objetivo e em especial a minha orientadora Jonice que me fortaleceu, auxiliou e confiou em mim durante toda essa trajetória. Por fim agradeço a todos os amigos que conquistei nessa jornada e que me ajudaram com conselhos e dicas e a todos os eventos que participei os quais me incentivaram a continuar e buscar sempre fazer o meu melhor.

# Resumo

NOLASCO, Diogo. **Identificação automática de áreas de pesquisa em C&T**. 2016. 82. Dissertação (Mestrado em Informática) – Instituto de Matemática, Instituto Tércio Pacciti, Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2016.

O crescimento da pesquisa, ciência e tecnologia a nível mundial e a consequente elevação da quantidade de informações armazenadas, como artigos e patentes, pelas bases de dados atuais gera dificuldades na organização e no desenvolvimento de análises qualitativas sobre esses documentos, como prospecções tecnológicas e mapeamentos de linhas de pesquisa. Tais dificuldades, como os elevados custos e recursos necessários para realizar essas análises, aumentam cada vez mais a importância da automatização para se trabalhar com o grande volume de dados gerado. Entre as principais tarefas de uma análise científica em um ambiente com múltiplas fontes e formatos de dados está a identificação das áreas presentes no conjunto. Este trabalho apresenta uma técnica integrada para a identificação automática de áreas de pesquisa presentes em uma coleção de documentos e sua posterior representação através de rótulos para facilitar a compreensão do seu conteúdo. Assim, ao utilizar dados de diversas fontes e processar a informação textual nelas contida serão realizadas as tarefas de: 1) Identificar a quantidade de áreas presentes na coleção; 2) dividir os dados entre as áreas e 3) criar uma representação destas áreas que expresse seu tema principal. Em cada uma das etapas do processo serão avaliadas as opções existentes e adaptadas ou criadas novas alternativas para adequação ao cenário. Foram Realizadas avaliações qualitativas comparando os resultados com o trabalho manual realizado habitualmente para a mesma tarefa e para cada passo são comparadas as abordagens criadas ou adaptadas com outras já existentes. Os resultados mostram que conseguiu-se satisfatoriamente identificar as áreas da coleção e rotular as áreas de forma representativa e com boa similaridade ao trabalho humano. Novas oportunidades na pesquisa de mineração científica e em análises automáticas são abertas com esses resultados principalmente no uso dos variados tipos de dados e em análises temporais.

Palavras-chave: Mineração textual, Modelagem de tópicos, Bases científicas

# Abstract

NOLASCO, Diogo. Automatic Research Areas Identification in C&T. 2016. 82. Thesis (Master in Informatics) – Instituto de Matemática, Instituto Tércio Pacciti, Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2016.

The growth of research, science and technology at a global level and the consequent growing amount of stored information, such as articles and patents, in actual databases, lead to organization and qualitative analysis issues on these documents, as technological prospecting and mapping research areas. These issues, like growing costs and resources, make automatization even more important to work with the big volume of generated data as time passes. Among the main tasks of a scientific analysis in a multi-source and multi format environment is the detection of the research areas in a given collection. This work presents an integrated method for automatic identification of research areas in a document collection and posterior representation of its contents too, as a means to facilitate content comprehension. Then, by using data from different sources and processing them, we are going to: 1) Detect the number of research areas in the collection; 2) group the document data by area and 3) create a representation of these areas that transmits to the user the main subject of the research. In each of these steps, we assess the existing options and adapt or create new alternatives for the domain. We make experiments with qualitative evaluation comparing results with the manual work commonly used for the same task and in each step we compare created approaches with existing ones. The results show that we can satisfactorily identify and label research areas automatically in a representative way and in a similar way as manual work does. These evidences and results mainly in multi-source data and time series analysis open new opportunities in scientific mining and automatic analysis

Keywords: Text Mining, Topic Modeling, Scholar Data

# Comunicação

**O conhecimento gerado neste trabalho foi parcialmente disseminado através de:**

Publicações:

- NOLASCO, D. ; OLIVEIRA, J. .Detecting Knowledge Innovation through Automatic Topic Labeling on Scholar Data. Proceedings of 49th Hawaii International Conference on System Sciences, 2016. p. 358. Hawai, USA. (Qualis A1)
- NOLASCO, D. ; OLIVEIRA, J. .Modelagem de Tópicos e Criação de Rótulos: Identificando Temas em Dados Semi e Não-estruturados. Simpósio Brasileiro de Banco de Dados - Tópicos em gerenciamento de dados e informações, Eduardo Ogasawara, Vaninha Vieira. (Org.). 31ed.Porto Alegre: SBC, 2016, v. , p. 87-112. ISBN: 978-85-7669-344-4

Minicursos:

- "Modelagem de Tópicos e Criação de Rótulos: Identificando Temas em Dados Semi e Não-estruturados" no 31º Simpósio Brasileiro de Banco de Dados, 2016.
- "Modelagem de Tópicos e Criação de Rótulos: Identificando Temas em Dados Semi e Não-estruturados", Departamento de Ciência da Computação, Universidade Federal do Rio de Janeiro, 2016.

Palestra:

- "Mineração de Dados para Descoberta e Rastreio de Áreas Científico-Tecnológicas" no 2º Workshop Brasileiro da Ciência dos Dados, Tech Mining e Inovação, 2015.

# Lista de Equações

# Lista de Figuras

# Lista de Quadros

# Lista de Siglas

| | |
|---|---|
| C&T | Ciência & Tecnologia |
| GQM | Goal, Question, Metric |
| KDD | Knowledge Discovery and Data Mining Conference |
| KL | Kullback-Leibler |
| LSA | Latent Semantic Analysi*s* |
| LDA | Latent Dirichlet Allocation |
| P&D | Pesquisa e Desenvolvimento |
| PLSA | Probabilistic Latent Semantic Analysis |
| SDC | Scholar Data Challenge |
| TTM | Temporal Txet Mining |

# Sumário

# 1 Introdução

Neste capítulo serão apresentados a inspiração para este trabalho com diversos exemplos e dados, o foco da solução, com uma descrição mais detalhada do problema da identificação das áreas de pesquisa e os objetivos almejados com a proposta. Cada um desses temas é retratado nas próximas seções de: Motivação, Problema e Objetivos, respectivamente.

## 1.1 Motivação

Nas últimas décadas têm ocorrido um grande aumento na produção científica e tecnológica mundial, sobretudo no Brasil e no restante dos países emergentes que compõem o grupo denominado BRICS (que inclui também Rússia, China, Índia e África do Sul) (ABRIL, 2012).

No caso específico da Ciência, temos na Figura 1 o crescimento mundial na publicação de artigos que serve como um indicador do aumento nas pesquisas. Em relação ao Brasil podemos ver na Figura 2 o crescimento correspondente nas publicações e na Figura 3 (SOARES, 2014) a comparação do número de citações da produção brasileira em relação a alguns países desenvolvidos e outros emergentes. O número de citações serve para indicar que o país cresce não apenas na quantidade, mas também na qualidade e consequente reconhecimento da comunidade.

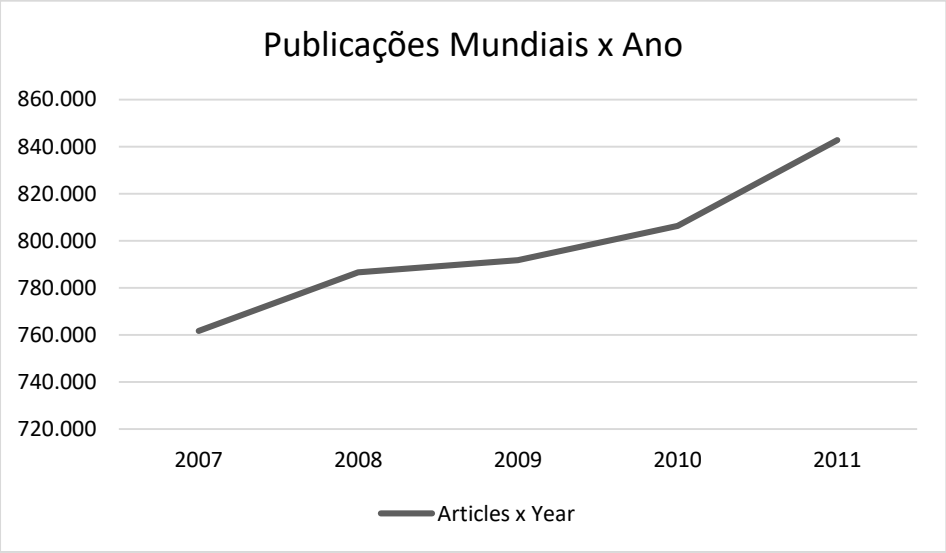



Figura 1: Número de publicações mundiais em periódicos no período 2007-2011. Fonte: Adaptado de (WORLD BANK, 2016)


Figura 2: Número de publicações brasileiras em periódicos no período 2008-2011. Fonte: Adaptado de (WORLD BANK, 2016).

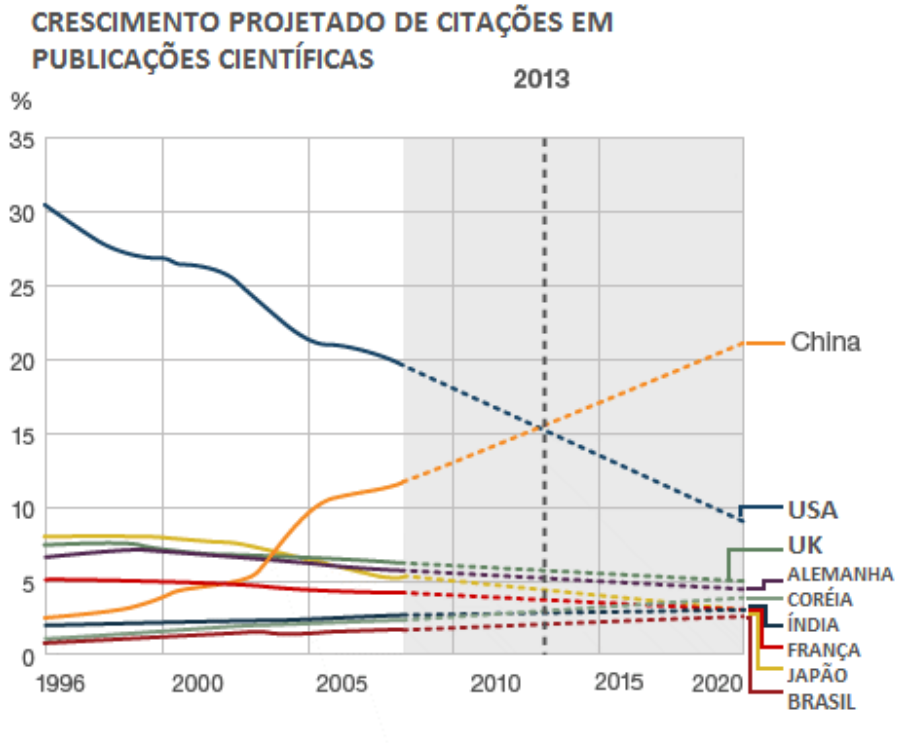



Figura 3: Crescimento de citações de alguns países em publicações científicas. Fonte: (SOARES, 2014).

Já no âmbito tecnológico, temos como principal parâmetro indicador o crescimento do número de patentes no mundo, que pode ser visto na Figura 4. O cenário nacional é proporcional e vem crescendo como mostra a Figura 5. Inclusive a produção de novas pesquisas e tecnologias tornou-se um ponto estratégico para o governo, que investe em projetos como a Rede Nacional de Ensino e Pesquisa e as parcerias com centros de inovação para o estímulo, desenvolvimento e proteção do conhecimento produzido (MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E INOVAÇÃO, 2016).

Nos países desenvolvidos as atividades de Ciência e Tecnologia (C&T) já são reconhecidas como componentes fundamentais para o desenvolvimento econômico, tecnológico e industrial das nações. Em discurso, o atual Presidente dos Estados Unidos, por exemplo, já ressaltou que a ciência, tecnologia, engenharia e medicina são críticas para a prosperidade de uma nação (THE NATIONAL ACADEMIES PRESS, 2016).

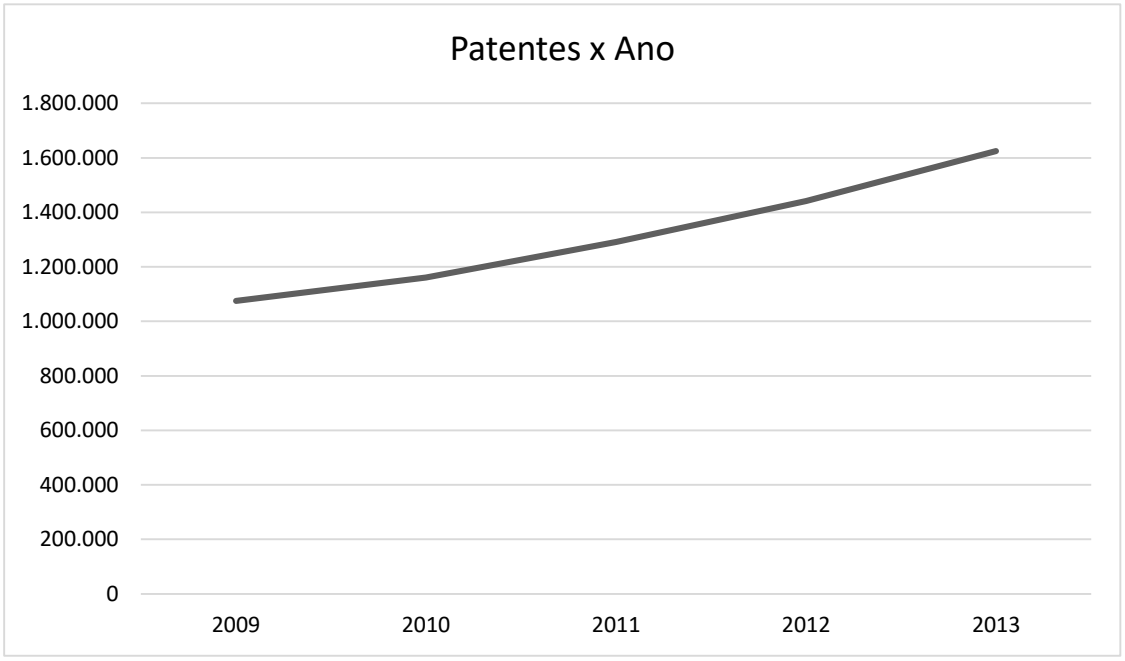



Figura 4: Número de patentes registradas em todo o mundo no período 2009 – 2013. Fonte: Adaptado de (WORLD INTELLECTUAL PROPERTY ORGANIZATION - WIPO, 2013)

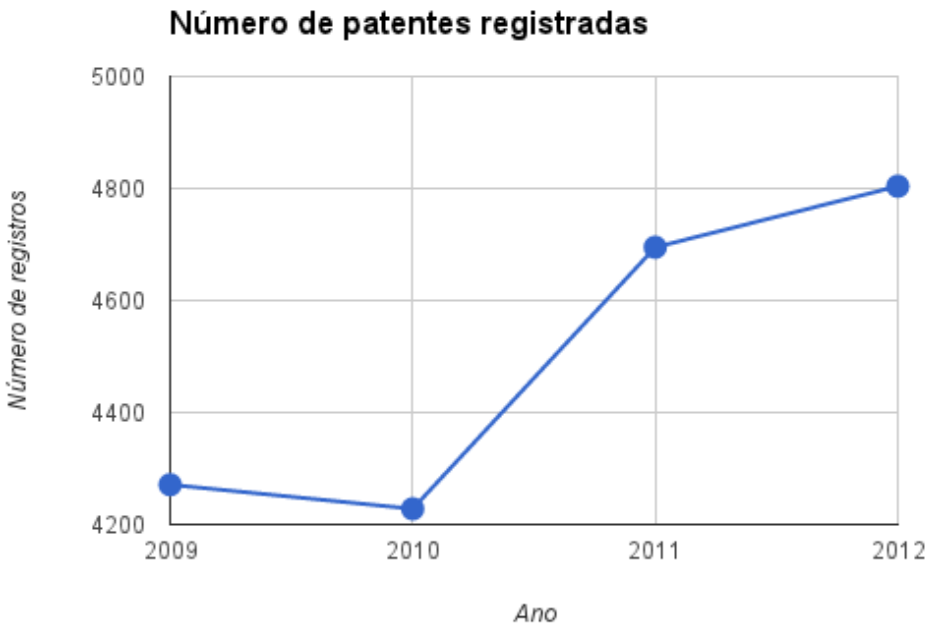



Figura 5: Número de patentes depositadas no Brasil no período 2009 – 2012.
Fonte: Adaptado de (WORLD INTELLECTUAL PROPERTY ORGANIZATION - WIPO, 2013)

No Brasil, o aumento dos esforços no sentido de apoio e estímulo a Ciência e Tecnologia vem acompanhado do reconhecimento dado pelas autoridades. O ex-ministro da Ciência, Tecnologia e Inovação, Aloisio Mercadante, em 2012 ressaltava a importância estratégica que o governo estava dando a área. Em declaração, enfatizou a importância da ciência, tecnologia e inovação ter sido colocada pelo governo e pelo ministério como eixo estruturante do desenvolvimento do Brasil dizendo que “Pela primeira vez está no Plano Plurianual como um dos marcos e objetivos estratégicos do país. E é um ministério que está pensando a nova economia brasileira” (REDE NACIONAL DE ENSINO E PESQUISA, 2016).

## 1.2 Problema

Todo esse crescimento nos leva à geração de uma grande quantidade de dados, de maneira que se torna intratável, cara e lenta a análise destes por especialistas. Por outro lado existe um empreendimento cada vez maior em se disponibilizar os dados e cada vez mais dados se tornam abertos, principalmente os governamentais.

Em contrapartida à sobrecarga de dados gerados, houve uma evolução nos últimos anos na maneira de se lidar com estes graças ao advento de novas técnicas automatizadas e de inteligência no campo da computação. Técnicas de Processamento de Linguagem Natural, Mineração de Textos e Recuperação de Informação, entre outras, já começam a ser utilizadas para extração de novas informações a partir de grandes massas de dados (MANYIKA et al., 2011). A própria evolução tecnológica está nos permitindo acompanhar o tratamento de grande volume de dados com recursos computacionais mais poderosos capazes de atender a demanda por suporte e análise.

Essa nova disponibilidade de muitos dados em formatos diversificados abre novos desafios e novas possibilidades na pesquisa sobre a pesquisa ("*R on R*", na sigla em inglês). Exemplos são a integração das informações oriundas do campo da pesquisa, da inovação e da sociedade, o rastreamento da evolução tecnológica e a análise da dinâmica dos grupos envolvidos.

Mesmo com estas oportunidades e com a possibilidade computacional para torná-las reais, o que se faz atualmente é muito limitado e faz pouco proveito dos avanços tecnológicos alcançados nas últimas décadas. As análises sobre dados de C&T ainda é feita de maneira semiautomática e muitas vezes manual por especialistas, o que torna todo o processo mais lento e dependente de pessoal. Além disso, os indicadores de desenvolvimento, como os bibliométricos focam exclusivamente na atividade científica ou na tecnológica.

Os próprios estudos atuais na área exibem essa necessidade. (TSENG et al., 2009) diz que o monitoramento das pesquisas em crescimento sempre foi de interesse para gestores políticos das áreas de ciência e tecnologia e que a criação de mecanismos automáticos facilitaria muito a tarefa. (SHIBATA; KAJIKAWA, 2009) afirma que para inovadores e pioneiros é essencial a detecção de áreas de pesquisa emergentes antes de seus competidores. Afirma também que para gestores de pesquisa e desenvolvimento (P&D) a identificação de novas áreas de pesquisa dentre um alto número de artigos acadêmicos se tornou uma tarefa significante. Inclusive, segundo os autores há uma demanda em se descobrir os domínios das pesquisas e os tópicos discutidos dentro deles. Por fim, (CHEN, 2006a) ressalta que a detecção do desenvolvimento das disciplinas na ciência pode aumentar significativamente a habilidade dos cientistas em lidar com mudanças e eventos inesperados.

Um exemplo dessas necessidades é a identificação e rastreio de áreas de pesquisa. Esta tarefa permite por exemplo que uma organização possa mapear as áreas de conhecimento com as quais trabalha, que um governo selecione as tecnologias mais relevantes para investimento ou que um pesquisador visualize as linhas de pesquisa mais interessantes para iniciar um trabalho ou cooperação. Todos esses exemplos dependem de uma organização dos dados e de uma certa classificação temática em um vasto universo de documentos. Em um mundo de dados é usualmente necessário para esta tarefa que se delimite o escopo de informações a uma base de dados e que estes sejam estruturados ou uniformes. A partir daí é preciso que um especialista veja as informações e descubra a que campo do conhecimento pertence um documento ou grupo de documentos. Se a análise for temporal, o processo deve ser repetido para cada conjunto de dados.

Assim, os resultados ficam limitados à base utilizada, ao conhecimento dos especialistas envolvidos e ao escopo definido. Isto se torna ainda mais grave quando se percebe que as áreas científico-tecnológicas vêm crescendo, se diversificando e se tornando cada vez mais interdisciplinares. Em conjunto vem o fato da maior diversidade de fontes de C&T, com o aumento da disponibilização da informação por meio de movimentos como a Ciência aberta (SOARES, 2014) e o uso de redes sociais *on-line*.

Além disso, a identificação de áreas tem um impacto em várias outras tarefas que são dependentes dessa. Um exemplo seria a detecção das instituições que possuem maior conhecimento em determinada linha de pesquisa ou dos principais profissionais que atuam nessa linha. Outra tarefa dependente da identificação seria o estudo do relacionamento entre tópicos de interesse dos pesquisadores e o próprio desenvolvimento desses tópicos ao longo do tempo. Uma visão temporal poderia ser capaz de mostrar as áreas que se encontram saturadas, se ramificando ou se unindo, assim como sua ascensão e declínio.

Então, em virtude deste cenário e das perspectivas futuras, as formas de se realizar análises como as descritas anteriormente se tornam ainda mais caras e limitadas. Hoje temos novas maneiras de facilitar e tratar este panorama. O uso de técnicas automatizadas pode substituir os processos manuais e com o bônus de abarcar também toda a dinâmica da disponibilidade, diversidade e quantidade de dados. Através da computação é possível tratar dados heterogêneos, integrar diversas fontes, rastrear crescimentos, agrupar dados, obter estatísticas, entre outras funções.

A partir desta visão é elaborada uma proposta que visa automatizar a identificação de áreas de ciência e tecnologia (C&T) e do seu desenvolvimento a partir da multiplicidade de fontes e de dados encontrados atualmente. Nas próximas seções serão detalhadas a proposta, a metodologia e o plano de trabalho respectivamente idealizados.

## 1.3 Objetivos

### 1.3.1 Objetivo Geral

O objetivo geral é a criação de uma técnica integrada para a identificação e representação automática das áreas de ciência e tecnologia (C&T) para a gestão estratégica no nível do pesquisador, dos investidores e do próprio País.

### 1.3.2 Objetivos Específicos

Para atingir esse objetivo será necessário realizar a identificação de áreas científico-tecnológicas para obtenção dos tópicos tratados na área, ramificações e interesses dentro da mesma. Dentro deste objetivo estão as tarefas de se identificar o espaço ocupado pela área e o tópico relacionado correspondente.

Estes objetivos específicos são necessários para o rastreio automático do desenvolvimento de um domínio do conhecimento e de sua difusão.

- Definir um processo para a identificação das áreas
- Levantar e comparar as técnicas existentes
- Adaptar técnicas existentes (quando utilizadas) ao escopo do trabalho
- Propor novas técnicas quando as existentes não forem suficientes para o fim desejado
- Levantar, comparar e propor formas de representação de áreas
- Avaliar a técnica integrada proposta com dados reais

O diferencial da presente pesquisa deve ser a elaboração de meios computacionais para a resolução de tarefas que até hoje são realizadas de maneira manual ou semiautomática. Por isto é importante lembrar também que todas essas ações devem ser realizadas por recursos informáticos automatizados para que o objetivo almejado seja plenamente alcançado.

O escopo deste trabalho é delimitado pela sua aplicação na área de ciência e tecnologia. Assim, para a identificação de áreas e grupos de pesquisas serão considerados dados relacionados a estas áreas, como patentes e publicações.

Outro fator de delimitação é a necessidade de uso de formas automatizadas para a realização das metas, não sendo necessárias intervenções humanas na análise da grande quantidade de dados.

A proposta vem atender a demandas das áreas de C&T e de pesquisa sobre pesquisa. Os dados mostrados na introdução por si já mostram a necessidade de novos meios para tratar a quantidade e diversidade de dados da área.

Os próprios estudos atuais na área exibem esta necessidade. (TSENG et al., 2009) diz que o monitoramento das pesquisas em crescimento sempre foi de interesse para gestores políticos das áreas de ciência e tecnologia e que a criação de mecanismos automáticos facilitaria muito a tarefa. (SHIBATA; KAJIKAWA, 2009) afirma que para inovadores e pioneiros é essencial a detecção de áreas de pesquisas emergentes antes de seus competidores. Afirma também que para gestores de pesquisa e desenvolvimento (P&D) a identificação de novas áreas de pesquisa dentre um alto número de artigos acadêmicos se tornou uma tarefa significante. Inclusive, segundo os autores há uma demanda em se descobrir os domínios das pesquisas e os tópicos discutidos dentro deles. Por fim, (CHEN, 2006a) ressalta que a detecção do desenvolvimento das disciplinas na ciência pode aumentar significativamente a habilidade dos cientistas em lidar com mudanças e eventos inesperados.

Entre as contribuições desta pesquisa encontram-se:

- Uma técnica integrada para a detecção de forma automática de áreas de pesquisa em C&T.
- Uma técnica para selecionar o número ótimo de áreas presentes numa coleção.
- Uma nova técnica de rotulagem escalável no tamanho da base.
- A análise temporal dos tópicos de pesquisa para rastreio da evolução das áreas ao longo do tempo através da criação de grafos de evolução temporal.

## 1.4 Nomenclatura Geral

Para uniformizar os termos utilizados para descrever as áreas de pesquisa, serão utilizados três termos distintos em diferentes partes do texto: Área, grupo e tópico.

O termo área será utilizado como sinônimo de área de pesquisa dado o escopo onde o trabalho é realizado. É usado em conjunto com os termos técnicos para facilitar a associação das técnicas com as aplicações ao escopo do trabalho.

O termo grupo será utilizado sobretudo na Seção 2 como um equivalente técnico de área. Nesta seção são utilizados termos específicos de técnicas de agrupamento que possuem uma nomenclatura específica na computação generalizando qualquer conjunto de dados agrupados como grupos.

O termo tópico será utilizado principalmente nas Seções 3, 4 e 5, como um equivalente conceitual de área. Nessas seções são discutidas abordagens que envolvem a modelagem de tópicos, a qual utiliza em suas definições formais o conceito de tópico como um equivalente do grupo nos agrupamentos tradicionais. Apesar disso, o grupo se refere ao conjunto de documentos que forma uma área de pesquisa e o tópico ao conjunto de termos que formam a mesma área. Pode-se dizer então que cada tópico possui um único grupo associado podendo utilizar ambos os termos intercambiando-os.

Além dos termos que são utilizados para as áreas de pesquisa vale apresentar mais dois termos específicos: token e palavras-chave. Ambos são utilizados principalmente nesta Seção e na Seção 4.

Um token na computação é um segmento de texto ou conjunto de caracteres que possui um significado e pode ser manipulado computacionalmente. Um exemplo simples seriam as palavras de um texto, onde cada uma pode ser um token pois possui significado próprio. Para datas por exemplo, já é possível a divisão em vários tokens distintos. A data “30/11/2016” pode ser um token com o signicado de data, dois tokens (“30/11” e “2016”) com signicado de “dia e mês” e “ano” ou três, um para o dia, mês e ano separadamente. Todas as possiblidades possuem significado próprio e a divisão aí depende do processo de *tokenização*. Esse processo simplesmente divide o texto em tokens de acordo com um padrão (para palavras, por exemplo, o divisor seria o espaço e para datas a “/”).

Uma palavra-chave seria uma palavra que resumiria o tema principal do texto no qual ela está contida. Um exemplo pode ser visto no próprio resumo desta dissertação. Uma frase-chave seria simplesmente a extensão deste conceito para frases.

O uso dos termos específicos de cada área tem por fim facilitar a compreensão de quem tem familiaridade ou pesquisa cada uma, porém essas definições de equivalências no texto ajudam a esclarecer os termos para todos os tipos de público.

## 1.5 Abordagem de solução

Para alcançar o objetivo desejado foi elaborada uma proposta que é composta de três partes principais: Agrupamento, Seleção de grupos e Representação dos grupos. A entrada ou matéria-prima para a identificação automática das áreas de pesquisa serão as coleções de documentos científicos (artigos, livros, patentes) e portanto uma etapa de tratamento ou pré-processamento pode estar presente dependendo da base utilizada. Então, para identificar e representar as áreas presentes nas fontes de dados, passa-se por um processo que engloba as três tarefas principais. Uma visão geral pode ser vista na Figura 6.

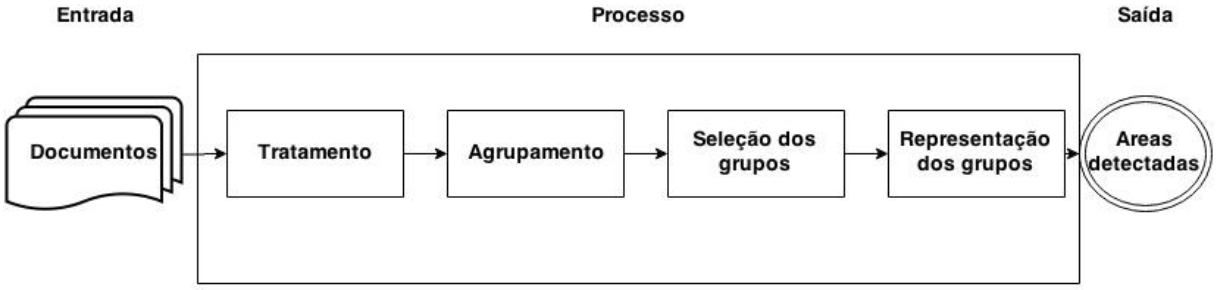



Figura 6: Esquema do processo de identificação de áreas de pesquisa

A entrada corresponde aos dados oriundos das fontes e homogeneizados para sua forma textual (textos em formatação). Algum tratamento pode ser necessário dependendo da fonte (por exemplo, remoção de formatação, *tokenização*).

Na fase de agrupamento os principais tópicos da coleção são extraídos e os documentos agrupados segundo o conjunto ao qual cada um pertence. Várias opções de números de tópicos são utilizadas para que se possa escolher o número ideal de tópicos presentes no corpus na próxima etapa.

Na fase de seleção dos grupos um algoritmo atribui uma pontuação para cada número de tópicos ou áreas testadas. O ranqueamento é feito baseado na análise de estabilidade dos tópicos de pesquisa, onde um número de tópicos que sofre grandes perturbações com pequenas mudanças recebe uma pontuação menor do que números de tópicos que se mantêm estáveis mesmo durante perturbações nos dados.

Depois das duas primeiras etapas obtém-se a coleção agrupada em conjuntos de documentos denominados grupos ou áreas (grupos e áreas são usados como sinônimos no cenário desta proposta). Para que estes possam ser usados em aplicações, visualizados por usuários ou para que sejam identificados automaticamente, uma última etapa de representação dos grupos é realizada através da rotulagem dos tópicos. Nesta etapa, os tópicos serão representados por palavras ou frases que explicitem o conteúdo temático de seus documentos.

Ao fim do processo, toda a coleção pode ser representada pelos grupos extraídos com seus respectivos tópicos atribuídos de maneira automática, desta forma minimizando intervenção humana no processo para que o torne mais fácil, rápido e menos custoso para as organizações.

Para cada etapa do processo foram realizados pesquisas e testes com as principais técnicas existentes. Quando não foi possível utilizar nenhuma delas satisfatoriamente no cenário tecnológico optou-se pela criação de novas técnicas ou adaptação e extensão de algoritmos ou métodos já existentes.

Assim, a proposta aqui descrita em uma visão geral é apresentada em maior profundidade nas seções que descrevem suas etapas.

As próximas seções abordam em maior detalhe as principais técnicas existentes para a realização das tarefas presentes em cada etapa, assim como a fundamentação teórica e as técnicas escolhidas ou criadas. Na Seção 2, são apresentados os conceitos, técnicas e algoritmos utilizados para o agrupamento dos documentos da coleção em áreas ou tópicos de pesquisa. Na Seção 3, o problema da seleção do número de áreas ótimo é exposto, o qual é um problema conhecido na pesquisa em agrupamento e presente em todos as técnicas analisadas. Terminando o processo, na Seção 4, a rotulagem de tópicos é apresentada como forma de representar as áreas de pesquisa com descrições concisas. Além de todas essas etapas, na Seção 5, é apresentada uma forma de utilizar a proposta de identificação de áreas para a análise temporal, ou seja, uma técnica para extrair as relações das áreas de pesquisa ao longo do tempo, seu desenvolvimento, ramificações, ascensão e declínio.

# 2 Agrupamento

Neste capítulo são apresentadas uma fundamentação teórica, algumas das principais técnicas de agrupamento existentes assim como a técnica que será utilizada na proposta. Esses temas estão respectivamente nas próximas seções de introdução, principais técnicas existentes e técnicas escolhidas

## 2.1 Introdução

O agrupamento, *clusterização* ou análise de agrupamentos é uma técnica de mineração de dados que a partir somente das informações das variáveis de cada item ou elemento, tem por objetivo agrupar automaticamente os dados de uma coleção em grupos geralmente disjuntos denominados *clusters* ou agrupamentos (HAN; PEI; KAMBER, 2011). É considerada uma técnica de aprendizado que geralmente envolve dois parâmetros básicos: N, um número de itens da base de dados (por exemplo, documentos) e K, o número de grupos (por exemplo, o número de áreas existentes).

Diferente do conceito de classificação (técnica de aprendizado supervisionado), o agrupamento é uma técnica mais “primitiva” onde não há nenhuma suposição a respeito dos grupos. Na classificação, existem classes predefinidas e através de um treinamento com exemplos de execução, os algoritmos “aprendem” como alocar os dados em cada classe, daí o nome aprendizado supervisionado. O agrupamento, ao contrário, não conhece de antemão as classes existentes e nem possui exemplos de como distribuir os dados entre os grupos, por isso realiza um aprendizado não-supervisionado.

A primeira publicação sobre um método de agrupamento foi feita em 1948, com o trabalho de (SORENSEN, 1948). Desde então muitos outros algoritmos de agrupamento já foram definidos. Qualquer método de agrupamento é definido por um algoritmo específico que determina como será feita a divisão dos N itens nos K grupos distintos e todos os métodos propostos são fundamentados na ideia de distância ou similaridade entre os agrupamentos, alocando os objetos em cada grupo segundo aquilo que cada elemento tem de similar em relação aos outros pertencentes ao mesmo grupo.

A ideia básica é que elementos que componham um mesmo grupo devem apresentar alta similaridade (isto é, sejam elementos bem parecidos, seguindo um padrão similar), mas

devem ter baixa similaridade em relação aos objetos de outros grupos. Dessa forma, todo agrupamento é feito com o objetivo de maximizar a homogeneidade dentro de cada grupo e maximizar a heterogeneidade entre grupos.

A grande vantagem do uso das técnicas de agrupamento é que, ao agrupar dados similares, pode-se obter de forma mais eficiente e eficaz as características de cada um dos grupos identificados. Isso fornece um maior entendimento da coleção de dados original, além de possibilitar o descobrimento de correlações interessantes entre os atributos dos dados que não seriam facilmente visualizadas sem o uso dessas técnicas.

O uso do agrupamento na presente proposta tem como objetivo realizar a divisão dos documentos em grupos de acordo com a sua temática. Como não se sabe a priori quantos assuntos distintos são abrangidos pela coleção, o uso do aprendizado não supervisionado permite dividir os documentos sem que se defina previamente as áreas ou classes que se busca. O algoritmo aqui deve ser capaz de alocar cada documento textual em sua área predominante e agrupá-los de maneira que as áreas de pesquisa sejam suficientemente distintas entre si enquanto uma mesma área possua similaridade de conteúdo entre seus elementos (serão usados os termos áreas e grupos como sinônimos no cenário deste trabalho).

## 2.2 Principais Técnicas Existentes

Dentre os inúmeros algoritmos de agrupamento existentes, podemos destacar três tipos usualmente utilizados para coleções textuais e que foram utilizados em testes empíricos iniciais para selecionar a técnica a ser utilizada: Agrupamento particional, Agrupamento hierárquico e Modelagem de tópicos.

### 2.2.1 Agrupamento Particional

Os algoritmos particionais dividem a base de dados em k grupos, onde o número k é dado pelo usuário. Esse é um ponto negativo do método pois esse domínio de conhecimento não é disponível para muitas aplicações, ou seja, raramente sabe-se de antemão quantos grupos ou áreas existem na coleção.

Inicialmente, o algoritmo escolhe k objetos como sendo os centros dos k grupos. Os objetos são divididos entre os k grupos de acordo com a medida de similaridade adotada, de modo que cada objeto fique no grupo que forneça o menor valor de distância entre o objeto

e o centro do mesmo. Primeiramente, atribui-se os elementos entre os grupos e então, o algoritmo utiliza uma estratégia iterativa para determinar quais objetos devem mudar de grupo, de forma que a função objetivo usada seja otimizada.

Após a divisão inicial, há duas possibilidades na escolha do elemento que vai representar o centro do grupo, e que será a referência para o cálculo da medida de similaridade. Pode-se utilizar a média dos objetos que pertencem ao grupo em questão, também chamada de centro de gravidade do grupo (esta é a abordagem conhecida como k-means) ou escolhe-se como representante o objeto que se encontra mais próximo ao centro de gravidade do grupo (abordagem conhecida como k-medoids), sendo o elemento mais próximo ao centro chamado de medóide.

O k-means é o mais popular e mais simples algoritmo particional e o escolhido para os testes iniciais realizados neste trabalho como representante dos algoritmos particionais. K-means foi descoberto independentemente por vários pesquisadores em campos de pesquisa diferentes (BALL; HALL, 1965; LLOYD, 1982; MACQUEEN, 1967) e mesmo tendo sido proposto há mais de 50 anos, ainda é um dos algoritmos mais utilizados para agrupamento devido à facilidade de implementação, simplicidade, eficiência e sucesso empírico e possui várias extensões desenvolvidas em várias plataformas.

A função objetivo mais utilizada nos métodos particionais é o erro quadrático, dado por:

$$E = \sum_{j=1}^{k} \sum_{x \in C_i} ||p - m_i||^2 , para\ k\ \in (1, n)$$

Equação 1: Erro quadrático como função objetivo

Na Equação 1, E é a soma do erro quadrado para todos os objetos na base de dados, p é o ponto no espaço representando um dado objeto, e $m_i$ é o representante do grupo $C_i$. Tanto p quanto $m_i$ são multidimensionais. Essa função objetivo dividida por n representa a distância média de cada objeto ao seu respectivo representante (ESTER et al., 1998). Os algoritmos terminam quando não existem atribuições possíveis capazes de melhorar esta função objetivo (COLE, 1998).

Os métodos particionais produzem agrupamentos simples. Esses algoritmos são efetivos se o número de grupos k puder ser estimado, se os grupos formados são convexos e

possuem tamanho e densidade similares (ANKERST et al., 1999). Esses métodos tentam fazer os k grupos tão compactos e separados quanto possível, e trabalham bem quando os grupos são compactos, densos e bastante separados uns dos outros, mas não são tão eficientes quando existem grandes diferenças nos tamanhos e geometrias dos diferentes grupos (GUHA; RASTOGI; SHIM, 1998). (HAN; PEI; KAMBER, 2011) observam que os mais conhecidos e usados métodos de particionamento são o k-means, o k-medoids, e suas variações.

Para coleções com dados de alta dimensionalidade (como normalmente é o caso de documentos textuais, onde cada termo ou *token* corresponde a uma dimensão), muitos dos métodos de agrupamento existentes não obtém bons resultados devido à maldição da dimensionalidade (BELLMAN, 2003), que torna as funções de distância problemáticas em espaços de alta dimensão.

Um exemplo de agrupamento particional pode ser visto na Figura 7, que mostra um agrupamento de documentos de acordo com a similaridade de suas dimensões (termos ou *tokens*).

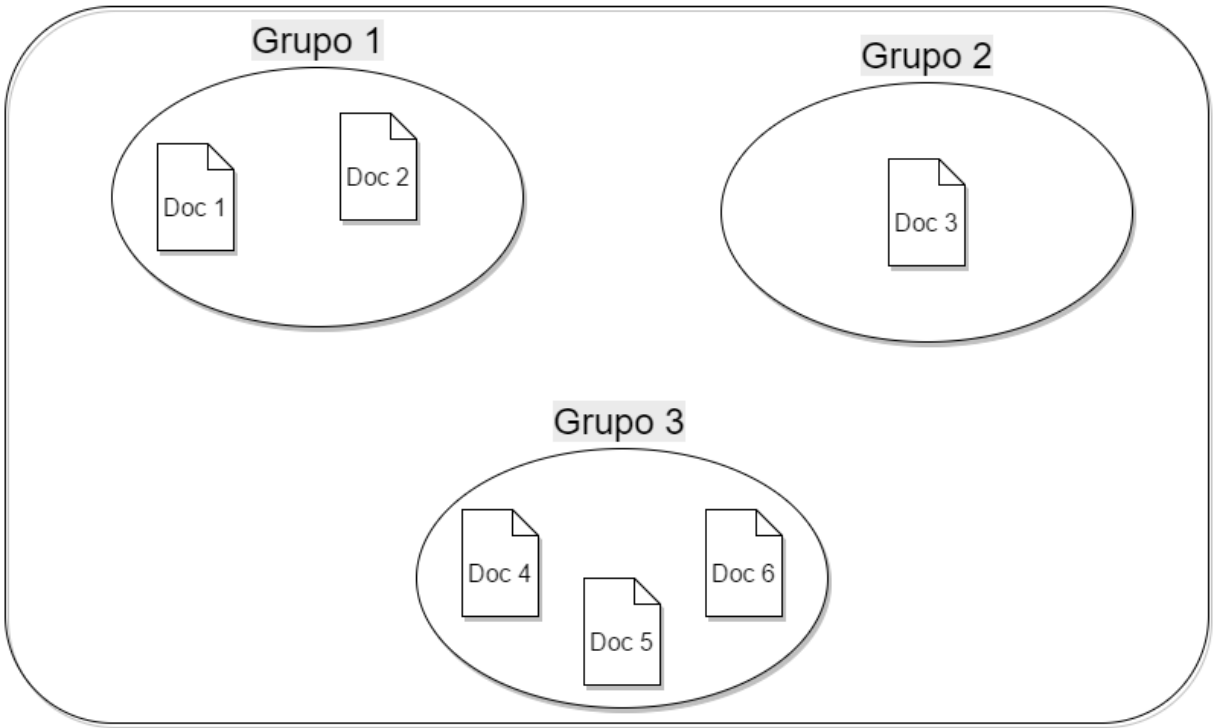



Figura 7: Exemplo de agrupamento particional.

### 2.2.2 Agrupamento Hierárquico

Algoritmos de agrupamento hierárquico organizam um conjunto de dados em uma estrutura hierárquica de acordo com a similaridade entre os elementos. Os resultados de um algoritmo hierárquico são normalmente mostrados como uma árvore binária ou dendograma, que é uma árvore que iterativamente divide a base de dados em subconjuntos menores. A raiz do dendograma representa o conjunto de dados inteiro (a coleção) e os nós folhas representam os indivíduos (no caso, os documentos). Um exemplo pode ser visto na Figura 8.

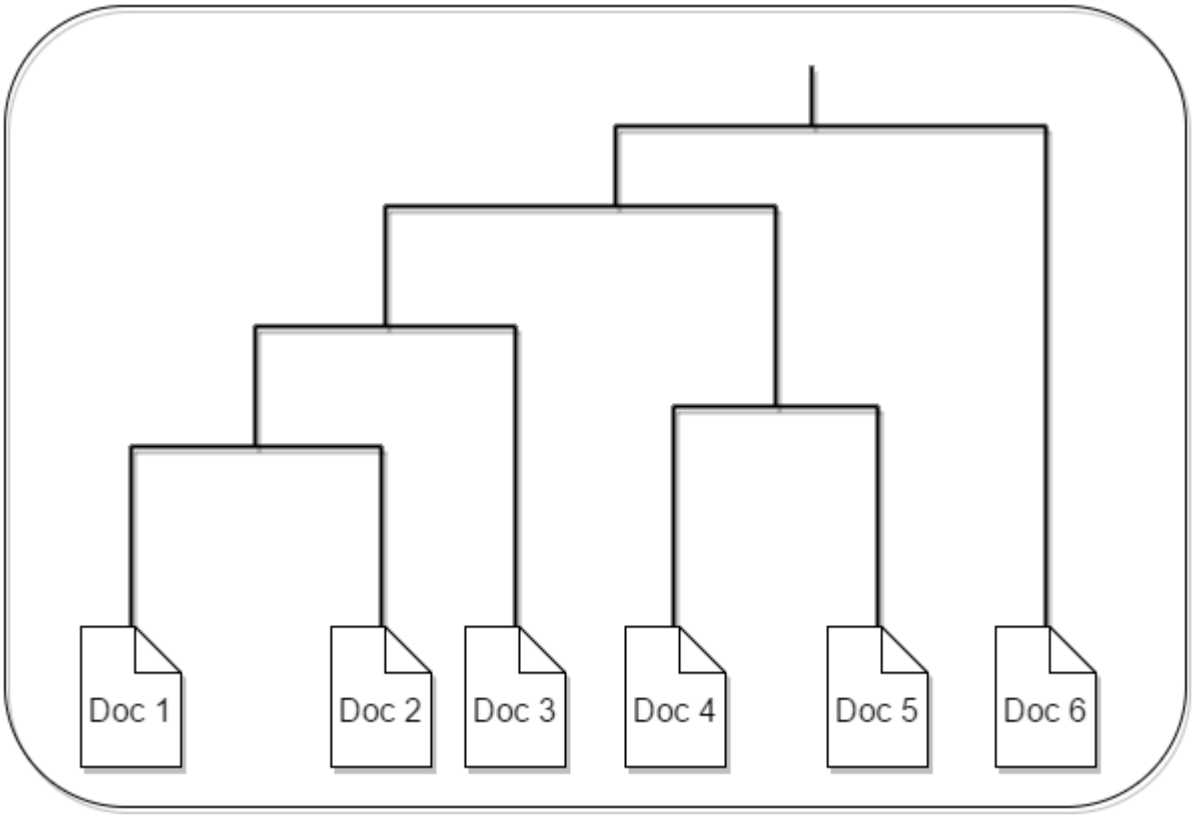



Figura 8: Esquema de um agrupamento hierárquico em uma coleção de documentos

O resultado de um agrupamento hierárquico pode ser obtido cortando-se o dendograma em diferentes níveis de acordo com o número de grupos k desejado. Esta forma de representação fornece descrições informativas e visualização para as estruturas de grupos em potencial, especialmente quando há realmente relações hierárquicas nos dados como, por exemplo, dados sobre evolução de espécies. Em tais hierarquias, cada nó da árvore representa um grupo da base de dados.

Importante lembrar, como foi visto na Figura 8, que na terminologia usada na Ciência da Computação, as folhas de uma árvore são os elementos individuais, no caso do exemplo os documentos. A raiz é o conjunto total que fica na parte de cima e agrega todos os elementos.

Para efeitos de visualização, seria como uma árvore real invertida, com a raiz em cima e as folhas para baixo.

O dendograma pode ser criado de duas formas:

1. Abordagem aglomerativa (*bottom-up*): Parte-se das folhas para a raiz. Inicia-se considerando cada objeto como sendo um grupo, totalizando n grupos (sendo n o número de elementos a serem agrupados). Em cada etapa, calcula-se a distância entre cada par de grupos. Estas distâncias são geralmente, armazenadas em uma matriz de dissimilaridade simétrica. Então, escolhe-se 2 grupos com a distância mínima e junta-os. A seguir, atualiza-se a matriz de distâncias. Este processo continua até que todos os objetos estejam em um único grupo (o nível mais alto da hierarquia), ou até que uma condição de término ocorra (AGRAWAL et al., 1998; HAN; PEI; KAMBER, 2011; R. NG, 1994).
2. Abordagem divisiva (*top-down*): Parte-se da raiz para as folhas. Nesta abordagem o processo é o inverso da abordagem *bottom-up* por começar com todos os objetos em um único grupo. Em cada etapa, um grupo é escolhido e dividido em dois grupos menores. Este processo continua até que se tenham n grupos (número total de elementos possíveis) ou até que uma condição de término, por exemplo, o número de grupos k desejado aconteça.

Os métodos aglomerativos são mais populares do que os métodos divisivos. (ZHANG; RAMAKRISHNAN; LIVNY, 1996) dizem que os métodos hierárquicos não tentam encontrar os melhores grupos, mas manter junto o par mais próximo (ou separar o par mais distante) de objetos para formar grupos. Também salientam que a melhor estimativa para a complexidade de um algoritmo prático por método hierárquico é $O(n^2)$ o que o torna ineficiente para valores de n grandes.

Neste trabalho a abordagem utilizada para se testar essa técnica foi a aglomerativa por ser a mais utilizada e comum.

### 2.2.3 Modelagem de tópicos

A modelagem de tópicos, apesar de ser um método estatístico para descobrir temas na estrutura de um corpus, também é vista como um agrupamento *fuzzy* ou *soft* (OLIVEIRA; PEDRYCZ, 2007). Como visto anteriormente, o agrupamento em geral divide os dados em grupos baseado em suas informações ou dimensões. Consiste em dois parâmetros básicos: N,

um número de casos da base de dados (por exemplo, documentos) e K, o número de grupos (por exemplo, o número de temas existentes).

Os tópicos extraídos pela modelagem podem ser então vistos como os grupos e os dados agrupados como os casos. Além disso, pode-se dividir as técnicas de agrupamento em dois tipos principais: *hard clustering* e *soft clustering* (ARABIE; HUBERT, 1996). O primeiro é o mais usual onde cada caso é associado a um e somente um grupo. Já o último, onde se encaixa a modelagem de tópicos, pode atribuir a cada caso um ou mais grupos com diferentes proporções (que no caso da modelagem é representado pela probabilidade de cada grupo).

Assim, a modelagem probabilística de tópicos é uma abordagem para atacar o problema do agrupamento e organização de dados, principalmente de conteúdo textual e cujo objetivo principal é a descoberta de tópicos e a anotação de grandes coleções de documentos por classificação temática. Tais métodos analisam quantitativamente as palavras dos textos originais para descobrir os temas presentes nos mesmos. Os algoritmos de modelagem de tópicos não requerem nenhum conhecimento prévio dos elementos e os tópicos emergem da análise dos textos originais (BLEI, 2012).

O campo de pesquisa em modelagem de tópicos a partir de documentos de textos teve inicialmente um marco com o desenvolvimento da técnica conhecida como Análise de semântica Latente (*Latent Semantic Analysis* ou LSA) (LANDAUER; DUMAIS, 1997). No LSA, utilizou-se do ferramental da álgebra linear para decompor um corpus nos seus temas constituintes, mais especificamente através da aplicação da decomposição SVD (*Singular value decomposition*) numa matriz com a contagem de frequência dos termos ao longo dos documentos de uma coleção. Na área de pesquisa em recuperação de informações, o LSA é utilizado para retornar documentos correspondentes a partir de uma busca por palavras-chave, categorizar documentos e generalizar resultados através de documentos equivalentes em diversas línguas (CHANG et al., 2009). Na modelagem de tópicos, o modelo LDA (Alocação Latente de Dirichlet, do inglês *Latent Dirichlet Allocation*)(BLEI, 2012) é um dos mais populares e serviu como base para a criação de muitos outros modelos probabilísticos. As fundações do modelo LDA foram baseadas no LSA e PLSI (*Probabilistic Latent Semantic Indexing*, uma evolução do LSA com o uso de fórmulas probabilísticas (BLEI; LAFFERTY, 2009; STEYVERS; GRIFFITHS, 2007).

No caso dos algoritmos de modelagem de tópicos a abordagem aqui baseia-se em criar uma distribuição de grupos para cada termo de um documento textual e uma distribuição de grupos para cada documento. Baseado nessas distribuições pode-se agrupar os documentos de acordo com as probabilidades associadas a cada grupo. Um exemplo deste tipo de técnica é ilustrado na Figura 9, que mostra quatro documentos associados aos tópicos “Genética”, “Evolução”, “Doenças” e “Computadores” (A largura das arestas que conectam os documentos aos grupos indicam a proporção do tópico presente no documento).

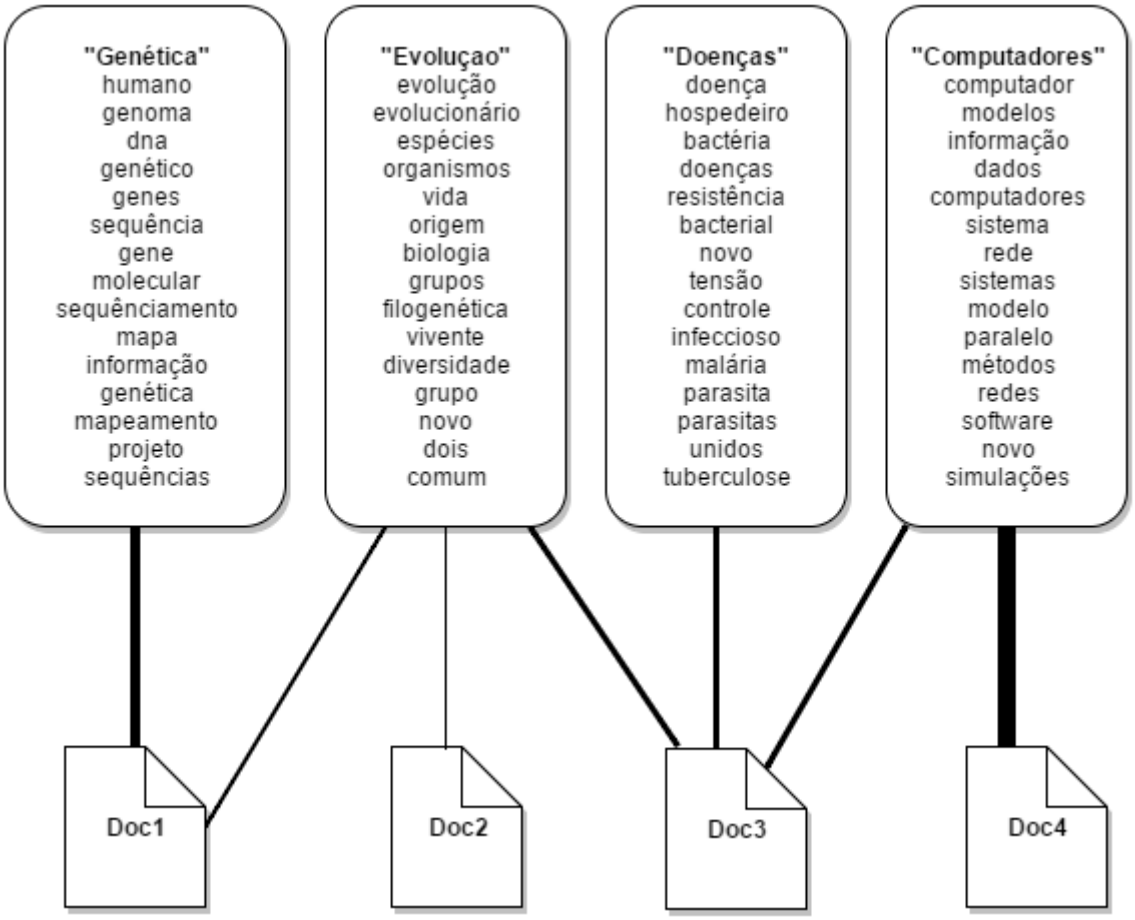



Figura 9: Exemplo de associação entre tópicos e documentos

Diferentemente dos agrupamentos particional e hierárquico que produzem grupos simples (cada elemento pertence a um grupo), a modelagem de tópicos realiza um agrupamento “leve” (*soft clustering*). Neste tipo de agrupamento, cada documento tem uma maior proporção de determinado tópico porém também pode conter outros tópicos em menores proporções. Assim, um documento por exemplo pode estar relacionado com a área

de “agrupamento” e mesmo assim possuir uma certa relação com outra área como “aprendizado de máquina” em menor proporção.

O algoritmo utilizado para esta abordagem nesta pesquisa foi o LDA que se apresenta como o mais usado e exitoso representante dos algoritmos de modelagem de tópicos, principalmente para bases de dados textuais.

## 2.3 Técnicas Escolhidas

Baseado em testes empíricos e nos pontos positivos e negativos de cada técnica apresentada, optou-se pelo uso do algoritmo LDA como forma de agrupamento. Por criar distribuições de probabilidades ao invés de calcular distâncias, a modelagem de tópicos não fica sujeita à maldição da dimensionalidade. Fora isso, devido ao fato de realizar um agrupamento “leve”, a modelagem tende a apresentar resultados mais realistas (um documento por exemplo, pode estar relacionado com dois grupos em diferentes intensidades).

A Alocação Latente de Dirichlet (LDA) e outros modelos de tópicos fazem parte do campo de pesquisa mais amplo de modelagem probabilística. Nesse tipo de modelagem, os dados são tratados como oriundos de um processo generativo que contém variáveis ocultas. Esse processo define uma distribuição de probabilidade conjunta sobre as variáveis aleatórias observadas e as ocultas, a qual é usada para computar a distribuição condicional das variáveis ocultas dadas as variáveis observadas. Essa distribuição condicional também é chamada de distribuição posterior ou simplesmente “posterior”. As variáveis observadas são as palavras nos documentos e as variáveis ocultas são a estrutura de tópicos (como mostra a Figura 10). O problema computacional de inferir a estrutura de tópicos oculta a partir de um conjunto de documentos é o problema de computar a distribuição posterior – a distribuição condicional das variáveis ocultas dados os documentos (BLEI, 2012).

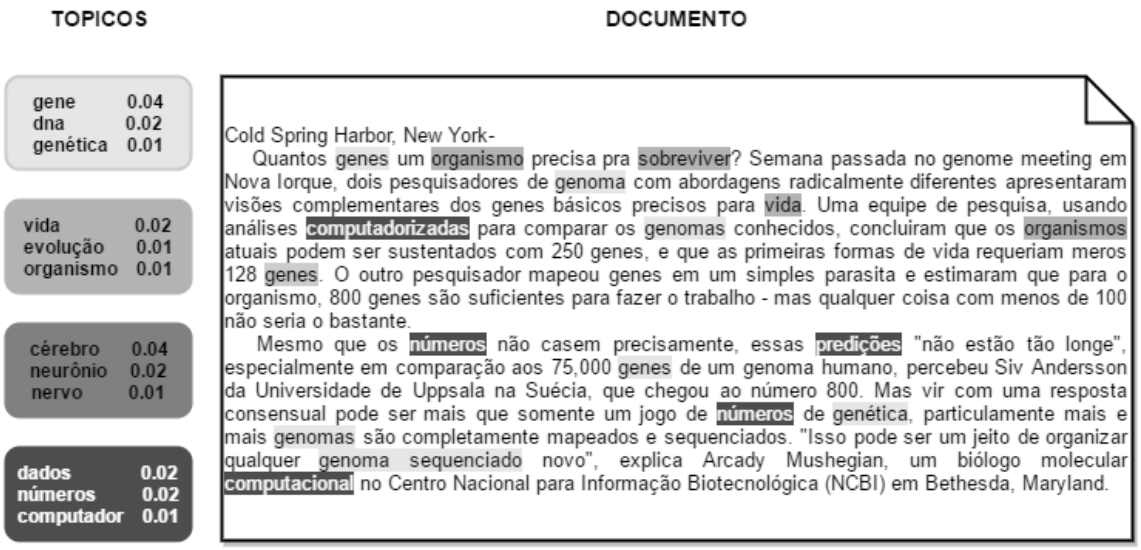



Figura 10: Relação entre os termos contidos no documento e os tópicos existentes (A cor de fundo representa a ligação da palavra ao tópico correspondente)

O processo generativo em LDA produz documentos de texto e os dados manipulados são as palavras ou termos que irão formar esses documentos. Trata-se de um processo imaginário, a partir do qual a estrutura de tópicos de uma coleção é obtida por inferência a partir da inversão daquele processo. Tecnicamente, o modelo assume que os tópicos são gerados antes dos documentos. Um tópico é definido como uma distribuição de probabilidade sobre um vocabulário fixo. Como exemplo, um tópico sobre genética será aquele que contém palavras relacionadas à genética com maior probabilidade de ocorrência. Em contraposição, um tópico que se relacione com qualquer outro assunto distinto conterá palavras sobre genética com probabilidade de ocorrência muito baixa ou zero. Todos os tópicos contêm distribuições com probabilidades sobre todo o vocabulário fixo, mas essas probabilidades só assumirão valores mais altos nos termos que caracterizam o tópico.

O processo que gera os documentos em LDA é realizado em duas etapas. Para a geração de cada documento da coleção, tem-se que:

1. Uma distribuição sobre tópicos é escolhida aleatoriamente. Exemplo: num modelo com apenas 3 tópicos, uma distribuição sobre tópicos possível para um documento A pode exibir probabilidades 0.1, 0 e 0.9 de ocorrência dos tópicos x, y e z respectivamente.
2. Para cada palavra no documento:

a. Um tópico é escolhido aleatoriamente a partir da distribuição obtida no passo 1.
b. Uma palavra é escolhida aleatoriamente a partir do tópico (o qual é uma distribuição de probabilidade sobre o vocabulário) obtido em 2a.

Cada documento exibe tópicos em proporções distintas (passo 1), cada palavra em cada documento é obtida a partir de um dos tópicos (passo 2b), o qual por sua vez é escolhido a partir da distribuição sobre tópicos de um documento em particular (passo 2a). Esse modelo estatístico reflete a intuição de que documentos exibem múltiplos tópicos, um pressuposto que está por trás da formulação do modelo LDA.

O modelo LDA também pode ser descrito mais formalmente através da seguinte notação:

**1.** Dado os tópicos $\theta_{1:N}$, onde cada $\theta_n$ é uma distribuição sobre o vocabulário V.

**2.** As proporções dos tópicos para o d-ésimo documento são $\rho_d$, onde $\rho_{d,n}$ é a proporção do tópico n no documento d.

**3.** As atribuições de tópicos para o d-ésimo documento são $z_d$, onde $z_{d,i}$ é a atribuição do tópico para a i-ésima palavra no documento d.

**4.** Finalmente, as palavras observadas para o documento d são $w_d$, onde $w_{d,i}$ é a i-ésima palavra no documento d, a qual é um elemento do vocabulário V.

Com essa notação, o processo generativo em LDA corresponde à distribuição conjunta das variáveis observadas e ocultas representada pela expressão:

$$p(\theta_{1:N}, \rho_{1:D}, z_{1:D}, w_{1:D}) = \prod_{j=1}^{N} p(\theta_d) \prod_{d=1}^{D} p(\rho_d) \left( \prod_{i=1}^{I} p(z_{d,i}|\rho_d) p(w_{d,i}|\theta_{1:N}, z_{d,i}) \right)$$

Equação 2: Modelo de como os documentos são gerados

Sabendo como os documentos são gerados, pode-se fazer o processo inverso: dados os documentos (termos = variáveis observáveis), descobrir como foram gerados (tópicos = variáveis ocultas). A forma mais utilizada neste caso é através do algoritmo *Gibbs Sampling* (CASELLA; GEORGE, 1992) que, simplificadamente consiste em:

1. Para cada documento d distribuir cada palavra w entre $\theta$ tópicos (criar uma distribuição inicial)
    a. Para cada documento d:
        i. Para cada palavra w de d:
            1. Para cada tópico $\theta$ computar p($\theta$|d) e p(w|$\theta$)
            2. Atribuir um novo tópico para w com p($\theta$|d)*p(w|$\theta$)

Onde p($\theta$|d) é a proporção de palavras no documento d que estão associadas ao tópico $\theta$ e p(w|$\theta$) é a proporção de associações de todos os documentos ao tópico $\theta$ oriundas da palavra w. Já p($\theta$|d)*p(w|$\theta$) é essencialmente a probabilidade do tópico $\theta$ ter gerado a palavra w. Esse processo se repete até que se atinja um estado de equilíbrio na distribuição de probabilidades.

Ainda que a modelagem LDA consiga realizar uma divisão automática de coleções de milhares de documentos, o que não seria possível alcançar por anotação humana, é preciso cautela no uso e na interpretação dos resultados obtidos a partir desse modelo. Os tópicos e sua distribuição ao longo dos documentos obtidos a partir da modelagem LDA assim como o de outros modelos de extração de tópicos não são “definitivos”. Uma modelagem de tópicos aplicada em uma coleção sempre irá produzir padrões a partir do corpus, ainda que os mesmos não estejam “naturalmente” presentes na coleção. Portanto, o LDA deve ser visto como uma ferramenta para a exploração de dados, que aliada a outras técnicas pode ser aplicada a diversos problemas e onde os tópicos representam um resumo do corpus que seria impossível de obter manualmente. De qualquer forma, sob essa perspectiva exploratória a análise de um modelo de tópicos pode revelar conexões entre documentos e no interior dos mesmos que não seriam óbvias a olho nu e pode ainda encontrar coocorrências inesperadas entre termos (BLEI; LAFFERTY, 2009).

A superação das limitações do modelo LDA é uma área de pesquisa ativa e duas abordagens importantes no desenvolvimento de novos modelos de tópicos podem ser destacadas: a criação de novos modelos através do relaxamento de alguns pressupostos assumidos no LDA e a incorporação de metadados do corpus para enriquecer a modelagem dos documentos. Descrevendo com mais detalhes pressupostos que motivaram o desenvolvimento de modelos estendidos de tópicos, um primeiro pressuposto assumido na

modelagem LDA está relacionado ao conceito de "*bag of words*", o qual parte do princípio de que a ordem das palavras num documento não é relevante. Apesar de não ser um pressuposto realista, ele é razoável se o único objetivo da aplicação for revelar a estrutura semântica de textos. Um segundo pressuposto assumido em LDA é o de que a ordem dos documentos não importa. Esse pressuposto pode não ser realista ao se analisar coleções que atravessam anos ou séculos, pois nesses casos é importante considerar que existem alterações nos tópicos ao longo do tempo. O terceiro pressuposto em destaque é o de que o número de tópicos é conhecido e fixo. Na modelagem LDA, um dos parâmetros que deve ser definido a priori é justamente o número de tópicos a serem extraídos. O quarto e último pressuposto considerado é o de que os tópicos são independentes, o que impede de modelar a correlação entre os mesmos (BLEI, 2012). Outra área de pesquisa que estende e aprimora os resultados obtidos com o LDA é a do desenvolvimento de técnicas como a rotulagem, análise temporal e de correlação entre tópicos que trabalham em cima dos resultados do algoritmo sem a necessidade do desenvolvimento de novos programas ou de novas execuções para aplicação das respectivas técnicas.

Quadro 1: Comparativo das técnicas de agrupamento

| **Técnica** | **Características Principais** | **Vantagens** | **Desvantagens** |
|---|---|---|---|
| Agrupamento Particional | Cria grupos disjuntos dos elementos | Mais tradicional e utilizado para diversos tipos de dados | Exige conhecimento dos grupos para definição de parâmetros; pouco eficiente para dados com muitas dimensões |
| Agrupamento Hierárquico | Cria uma árvore ou dendograma dos elementos | Evita a definição prévia do número de grupos utilizando o dendograma; cria diversos níveis de agrupamento | Necessita de conhecimento da coleção para realizar o corte no dendograma formando os grupos; pouco eficiente para coleções grandes |

| | | | |
|---|---|---|---|
| Modelagem de Tópicos | Cria uma distribuição de probabilidades sobre os termos para cada tópico distinto | Cria uma distribuição de probabilidades evitando o problema de agrupar muitas dimensões; muito utilizado em coleções de textuais | Exige conhecimento prévio para definição do número de tópicos; Necessita de um agrupamento posterior dos documentos pois cada um é representado por uma mistura de tópicos |

# 3 Seleção do Número de Áreas

Neste capítulo é apresentado o problema da seleção do número de áreas ou grupos quando da utilização de um algoritmo de agrupamento, no caso, da modelagem de tópicos. As seções de introdução, principais técnicas existentes e técnicas escolhidas mostram o problema, as soluções mais comuns e a solução usada respectivamente.

## 3.1 Introdução

Na modelagem de tópicos em geral, um tópico em um coleção textual pode ser visualizado como uma distribuição de probabilidade ao longo dos termos presentes no corpus ou um grupo que define pesos para esses termos (WANG et al., 2012). A maioria das pesquisas sobre modelagem de tópicos têm focado no uso dos métodos probabilísticos tais como o LDA (BLEI; NG; JORDAN, 2003), utilizado aqui e outros, como por exemplo a análise probabilística de semântica latente (PLSA)(HOFMANN, 1999). Independentemente do algoritmo utilizado, uma consideração chave na aplicação da modelagem de tópicos é a seleção de um número apropriado de tópicos K para o corpus considerado. A escolha de um valor de k muito baixo irá gerar tópicos que são excessivamente amplos, enquanto escolher um valor que é muito alto irá resultar em um excessivo agrupamento dos dados. Para algumas coleções, é possível existir temas coerentes em vários níveis diferentes, desde grãos maiores até os menores, refletidos por vários valores de k apropriados.

Quando um resultado de um agrupamento é gerado utilizando um algoritmo que contém um elemento aleatório ou exige a seleção de um ou mais valores dos parâmetros, é importante considerar se a solução encontrada constitui uma solução "definitiva", que pode ser facilmente replicada. Técnicas de validação de grupos com base neste conceito foram mostradas para ajudar a escolher um número apropriado de grupos nos dados (LANGE et al., 2004; LEVINE; DOMANY, 2001). Neste contexto, a estabilidade de um modelo de agrupamento se refere à sua capacidade de se replicar de forma consistente, ou seja, obter soluções semelhantes em dados provenientes da mesma fonte.

Na prática, isto envolve repetir a fase de agrupamento utilizando diferentes condições iniciais ou aplicar o algoritmo para diferentes amostras do conjunto de dados completo. Um alto nível de concordância entre os agrupamentos resultantes indica alta estabilidade, por sua

vez, sugerindo que o atual modelo é apropriado para os dados. Em contraste, um baixo nível de concordância indica que o modelo é um ajuste ruim para os dados. A análise de estabilidade tem sido frequentemente utilizada em diversas aplicações (BERTONI; VALENTINI, 2005). O foco tem sido a seleção de modelos para as abordagens de agrupamento particional clássicas, como k-means (BEN-DAVID; PÁL; SIMON, 2007; LANGE et al., 2004) e de agrupamento hierárquico aglomerativo (BERTONI; VALENTINI, 2005; LEVINE; DOMANY, 2001).

## 3.2 Principais Técnicas Existentes

Uma variedade de métodos baseados no conceito de análise de estabilidade foi proposta para a tarefa de seleção do modelo. A estabilidade de um algoritmo de agrupamento se refere à sua capacidade de produzir consistentemente soluções semelhantes em dados provenientes da mesma fonte (BEN-DAVID; PÁL; SIMON, 2007; LANGE et al., 2004). Uma vez que apenas um único conjunto de itens de dados estará disponível em tarefas de aprendizagem não supervisionada, agrupamentos são gerados em perturbações dos dados originais. A principal aplicação da análise de estabilidade tem sido como uma abordagem robusta para selecionar os parâmetros do algoritmo (LAW, M. H.; JAIN, 2003), especificamente ao estimar o número ideal de grupos para um determinado conjunto de dados. Estes métodos são motivados pela observação de que, se o número de agrupamentos num modelo é muito grande, os agrupamentos realizados levarão a partições arbitrárias dos dados, resultando em soluções instáveis. Por outro lado, se o número de grupos é muito pequeno, o algoritmo de agrupamento vai ser obrigado a fundir subconjuntos de objetos que devem permanecer separados, também conduzindo a soluções instáveis. Em contraste, agrupamentos gerados usando um número ótimo de aglomerados serão geralmente consistentes, mesmo quando os dados são perturbados ou distorcidos.

A abordagem mais comum para a análise de estabilidade envolve perturbar aleatoriamente os dados por amostragem dos objetos originais para produzir um conjunto de sub-amostras para agrupamento, utilizando valores de k a partir de um intervalo pré-definido (LEVINE; DOMANY, 2001). A estabilidade do modelo de agrupamento para cada valor de k é avaliada usando uma medida de concordância para todos os pares de agrupamentos gerados em diferentes sub-amostras. Um ou mais valores de k são então recomendados, selecionando-os com base na maior pontuação de concordância.

(BRUNET; TAMAYO; GOLUB, 2004) propuseram uma abordagem baseada na estabilidade inicial para seleção de modelos de tópicos com base em atribuições de grupos separadas por itens em várias execuções do mesmo algoritmo usando diferentes inicializações aleatórias. Especificamente, para cada agrupamento realizado no mesmo conjunto de dados de n itens, uma matriz de conectividade n × n é construída, em que n(i, j) = 1 se as componentes I e J são atribuídas ao mesmo conjunto discreto, e (i , j) = 0 caso contrário. Repetindo este processo ao longo de τ execuções, uma matriz de consenso pode ser calculada como a média de todas as matrizes de conectividade τ. Cada entrada nesta matriz indica a fração de vezes que os dois itens foram agrupados juntos. Para medir a estabilidade de um determinado valor de K, um coeficiente de correlação é calculado em um agrupamento hierárquico da matriz de conectividade.

Nos trabalhos sobre LDA, (STEYVERS; GRIFFITHS, 2007) observaram a importância de identificar os temas que aparecem repetidamente em várias amostras de dados relacionadas, o que se assemelha ao conceito mais geral de análise de estabilidade (LEVINE; DOMANY, 2001). Os autores sugeriram comparar as duas execuções do LDA examinando uma matriz tópico x tópico construída a partir da distância simétrica Kullback-Liebler (KL) entre as distribuições de tópicos das duas execuções. Um trabalho visando medir a estabilidade de modelos de tópico via LDA foi descrito em (WAAL; BARNARD, 2008). Os autores propuseram uma abordagem centrada em documentos, onde os tópicos de duas execuções diferentes do LDA são combinados em um conjunto com base em correlações entre as linhas das duas matrizes de documentos de tópicos correspondentes. A saída foi representada como uma matriz de correlação documento-documento, onde a diagonal é estruturada pelos valores de correlação, indicando maior estabilidade. A este respeito, a abordagem é semelhante à abordagem de (BRUNET; TAMAYO; GOLUB, 2004).

Outras medidas de avaliação utilizadas para o LDA especificamente, incluem aquelas baseadas na coerência semântica dos principais termos derivados a partir de um único conjunto de tópicos, com respeito à coocorrência dentro do mesmo corpus ou em um corpus externo. Por exemplo, (NEWMAN et al., 2010) calculou correlações entre julgamentos humanos e um conjunto de medidas propostas calculadas, e descobriu que a Informação Pontual Mútua (PMI) alcançou o melhor ou quase o melhor resultado de todas as métricas consideradas. No entanto, essas medidas não foram utilizadas na seleção do modelo e seus

parâmetros e não consideram a robustez dos temas em várias execuções de um mesmo algoritmo.

## 3.3 Técnicas Escolhidas

Nesta seção é descrito um método geral baseado em estabilidade para selecionar o número de tópicos para a modelagem de tópicos utilizada no LDA. O método escolhido aqui é o apresentado por (GREENE; O'CALLAGHAN; CUNNINGHAM, 2014) e consiste na execução do agrupamento sobre amostras dos dados, mas utilizando a lista de termos característica da modelagem de tópicos para avaliar a estabilidade do modelo. Ao contrário dos métodos de análise de estabilidade não supervisionados discutidos anteriormente, no método escolhido o foco é o uso de recursos ou termos para avaliar a adequação de um modelo. Isto é motivado pela abordagem centrada em termos geralmente tomada na modelagem de tópicos, em que a prioridade é geralmente dada para a saída termo-tema e tópicos são resumidos usando um conjunto truncado dos termos mais relevantes (de maior probabilidade). Além disso, ao contrário da abordagem proposta em (BRUNET; TAMAYO; GOLUB, 2004), o método utilizado aqui não assume que os tópicos extraídos são separados e não requer o cálculo de uma matriz de conectividade ou a aplicação de um algoritmo de agrupamento subsequente.

Em primeiro lugar, nas próximas seções são detalhados: (i) uma métrica de similaridade para comparar duas listas de termos ranqueadas por relevância; (ii) uma medida de concordância entre duas modelagens de tópicos (execuções do agrupamento) quando representados por listas de termos por ordem de relevância; e (iii) Um método de análise de estabilidade para selecionar o número de tópicos em um corpus de texto.

### 3.3.1 Similaridade das listas de termos ranqueadas

Normalmente, a saída de um algoritmo de modelagem de tópicos se dá na forma de um conjunto de listas de termos por ordem de relevância contendo k listas (uma para cada tópico extraído), e denotada de agora em diante por $S = \{R_1, R_2, \dots, R_k\}$, onde cada $R_i$ é uma lista de termos. O tema do tópico $\theta_i$ produzido pelo algoritmo é representado pela lista $R_i$, contendo os principais termos que são mais característicos desse tópico de acordo com o critério de relevância. No caso do LDA este será composto dos termos com as maiores probabilidades na distribuição para cada tópico. Para algoritmos de agrupamentos

particionais ou hierárquicos, pode consistir nos termos com maior frequência em cada centroide do grupo, por exemplo.

Uma variedade de medidas simétricas pode ser usada para avaliar a semelhança entre um par de listas ranqueadas de termos ($R_i$, $R_j$). Uma abordagem simples seria empregar um método de sobreposição de conjuntos, como o índice de Jaccard (JACCARD, 1912). No entanto, tais medidas não levam em conta a informação da posição dos termos (tendo em vista que as listas aqui utilizadas são ranqueadas). Termos que ocorrem no topo de uma lista ordenada gerada por um algoritmo como o LDA irão, naturalmente, ser mais relevantes para um tópico do que aqueles que ocorrem na cauda da lista, que correspondem aos valores zero ou próximo de zero de relevância para o tópico em questão. Além disso, na prática, em vez de considerar todos os termos w em um corpus, os resultados da modelagem de tópico são apresentados usando a parte superior da lista, ou seja os t << m termos (os dez primeiros termos, por exemplo). Do mesmo modo, quando se mede a semelhança entre as listas ranqueadas, pode ser preferível considerar a listas truncadas com apenas t termos, para a economia de representação e reduzir o custo computacional da aplicação de várias operações de similaridade. No entanto, este, muitas vezes, pode levar a classificações indefinidas, onde os diferentes subconjuntos de termos estão sendo comparados.

Portanto, seguindo a medida de distância de listas ranqueadas proposta por (FAGIN; KUMAR; SIVAKUMAR, 2003), uma versão ponderada pelos termos mais relevantes do índice de Jaccard é proposta por (GREENE; O'CALLAGHAN; CUNNINGHAM, 2014) na técnica descrita aqui. Esta medida é adequada para o cálculo da similaridade entre pares de listas ranqueadas indefinidas, onde os elementos podem ser diferentes quando comparadas listas distintas. Especificamente, o trabalho define a métrica de média de Jaccard (AJ). Basicamente é calculada a média das pontuações de Jaccard entre cada par de subconjuntos dos termos mais bem classificados d em duas listas sendo comparadas, para a profundidade d ∈ [1, t]. Isto é:

$$AJ(R_i, R_J) = \frac{1}{t}\sum_{d=1}^{t} Yd\,(R_i, R_j)$$

Equação 3: Média de Jaccard.

Onde

$$Yd(R_i, R_j) = \frac{|R_{i,d} \cap R_{j,d}|}{|R_{i,d} \cup R_{j,d|}}$$

Equação 4: Índice de Jaccard.

de tal forma que $R_{i,d}$ é a cabeça da lista Ri até a profundidade d. Por exemplo, se são utilizados os dez primeiros termos para representar um tópico, a média é calculada entre os subconjuntos de um termo até dez termos. Esta é uma medida simétrica produzindo valores no intervalo [0, 1], em que os termos através de uma lista ordenada são ponderados de acordo com uma escala linear. Para demonstrar isso, um exemplo simples ilustrativo é mostrado no Quadro 2. Deve-se notar que, embora a pontuação Jaccard em profundidade d = 5 é relativamente alta (0,429), a pontuação média é muito mais baixa (0,154). Como a semelhança entre os termos ocorre no sentido das caudas das listas, estes termos têm menos peso do que aqueles no topo das listas, como "álbum" e "esporte".

Quadro 2: Exemplo da métrica da média de Jaccard aplicada em duas listas até a profundidade d = 5

| d | $R_{1,d}$ | $R_{2,d}$ | $Jac_d$ | AJ |
|---|---|---|---|---|
| 1 | Álbum | Esporte | 0,000 | 0,000 |
| 2 | Álbum, música | Esporte, melhor | 0,000 | 0,000 |
| 3 | Álbum, música, melhor | Esporte, melhor, vencedor | 0,200 | 0,067 |
| 4 | Álbum, música, melhor, prêmio | Esporte, melhor, vencedor, medalha | 0,143 | 0,086 |
| 5 | Álbum, música, melhor, prêmio, vencedor | Esporte, melhor, vencedor, medalha, prêmio | 0,429 | 0,154 |

### 3.3.2 Concordância entre Tópicos

Considerando agora o problema de medir a concordância entre duas modelagens de tópico k diferentes (duas execuções do LDA, por exemplo), representadas como dois conjuntos de listas ranqueadas $S_x = \{R_{x1}, R_{x2}, \dots, R_{xk}\}$ e $S_y = \{R_{y1}, R_{y2}, \dots, R_{yk}\}$, ambos contendo k listas. Constrói-se uma matriz M de similaridade k x k , de tal modo que uma entrada $M_{i,j}$ indica a concordância entre $R_{xi}$ e $R_{yj}$ (isto é, o i-ésimo tópico da primeira execução e o j-ésimo tópico da segunda), calculado com base na média de Jaccard. Em seguida, é encontrada a melhor correspondência entre as linhas e colunas de M (ou seja, as listas ordenadas nos conjuntos $S_x$ e $S_y$.

A permutação ótima π é encontrada usando o método húngaro (KUHN, 1955). A partir disso, pode-se produzir uma pontuação de concordância:

$$\text{concordância}(S_x, S_y) = \frac{1}{k}\sum_{i=1}^{k} AJ\ (R_{xi}, \pi(R_{xi}))$$

Equação 5: Concordância como somatório das médias de Jaccard

onde $\pi(R_{xi})$ denota a lista ranqueada em $S_y$ combinada com $R_{xi}$ pela permutação π (permutação ótima). Os valores sempre estão no intervalo [0, 1], onde uma comparação entre dois modelos de tópico com K grupos e idênticos (duas execuções do LDA com os mesmos tópicos, por exemplo) irá resultar numa pontuação de 1. Um exemplo ilustrando todo o processo de acordo com o que foi apresentado é mostrado na Figura 11.

Conjunto Ranqueado $S_1$:

$R_{11}$ = {esporte, ganhar, prêmio}

$R_{12}$ = {banco, finança, dinheiro}

$R_{13}$ = {música, álbum, banda}

| | $R_{21}$ | $R_{22}$ | $R_{23}$ |
|---|---|---|---|
| $R_{11}$ | 0.00 | 0.07 | 0.50 |
| $R_{12}$ | 0.50 | 0.00 | 0.07 |
| $R_{13}$ | 0.00 | 0.61 | 0.00 |

Conjunto Ranqueado $S_2$:

$R_{21}$ = {finança, banco, economia}

$R_{22}$ = {música, banda, prêmio}

$R_{23}$ = {ganhar, esporte, dinheiro}

$$\pi = (R_{11}, R_{23}), (R_{12}, R_{21}), (R_{13}, R_{22})$$

$$concordância(S_1, S_2) = \frac{0.50+0.50+0.61}{3} = 0.54$$

Figura 11: Exemplo mostrando como é realizada a comparação entre dois conjuntos de tópicos

### 3.3.3 Seleção do número de Tópicos

Com base na medida de concordância aqui apresentada e definida originalmente por (GREENE; O'CALLAGHAN; CUNNINGHAM, 2014), agora é apresentada uma extensão das medidas de concordância e de estabilidade definidas pelos autores para o nível do agrupamento e dos tópicos.

Para cada valor de k (número de tópicos) em um intervalo pré-definido $[k_{min}, k_{max}]$, procede-se como se segue. Primeiramente, é gerado um modelo de tópicos inicial sobre o conjunto completo de dados usando o algoritmo LDA escolhido para esta proposta, que

fornece um ponto de referência para a análise da estabilidade proporcionada usando k tópicos. Este será representado como um conjunto de listas ranqueadas de referência $S_0$, onde cada tema é representado pela lista ordenada dos seus principais t termos. Subsequentemente, τ amostras da coleção são construídas selecionando aleatoriamente um subconjunto de β × n documentos, onde 0 ≤ β ≤ 1 indica a relação de amostragem que controla o número de documentos de cada amostra (β = 1 seria a coleção inteira). Em seguida, são gerados τ modelos de tópicos com k tópicos cada, aplicando o algoritmo de modelagem de tópicos LDA para cada uma das amostras (executa-se o LDA em cada $\tau_i$), o que resulta em conjuntos de listas ranqueadas $\{S_1, \dots, S_\tau\}$, onde todos os tópicos também são representados usando os mesmos principais t termos do conjunto de referência (os t termos mais relevantes). Para medir a estabilidade global para k número de tópicos ou grupos, foi calculada a concordância média entre a lista ranqueada do conjunto de referência ($S_0$) e todos os outros conjuntos de classificação $\{S_1, \dots, S_\tau\}$ usando:

$$\text{estabilidade}(k) = \frac{1}{\tau}\sum_{i=1}^{\tau} concordância\ (S_0, S_i)$$

Equação 6: Estabilidade como o somatório das concordâncias entre modelagens

Este processo é repetido para cada número de grupos ou tópicos k ∈ $[k_{min}, k_{max}]$. Um resumo de todo o processo é dado na Figura 12.

Ao examinar as pontuações de estabilidade produzidas, um valor k final pode ser identificado com base nas pontuações mais altas. Também é possível plotar o resultado das pontuações e encontrar os picos no gráfico. A presença de mais de um pico indica que existem vários números de tópicos apropriados para o corpus em questão, o que é coerente com a existência de várias soluções alternativas em muitos problemas de análise de agrupamentos gerais (BAE; BAILEY, 2006). Um exemplo deste caso é mostrado na Figura 13 para um corpus de artigos jornalísticos adaptado de (GREENE; O'CALLAGHAN; CUNNINGHAM, 2014). Esta coleção possuía seis categorias ou grupos anotados manualmente, mas também é possível ver um pico em k = 3 nos valores de estabilidade, o que sugere que a estrutura temática existe em um nível mais alto também (por exemplo, categorias com maior abstração). Por outro lado, uma curva plana, sem picos, combinada com baixos valores de estabilidade, sugere fortemente que não existem tópicos coerentes no conjunto de dados.

Algoritmo:

1. Gerar $\tau$ amostras aleatórias da coleção, cada uma contendo $\beta \times n$ documentos.
2. Para cada número de tópicos $k \in [k_{min}, k_{max}]$:
    a. Aplicar o LDA na coleção de documentos completa contendo n documentos para extrair k tópicos e representar a saída como um conjunto de listas ranqueadas de referência $S_o$.
    b. Para cada amostra $\tau_i$:
        i. Aplicar o LDA em $\tau_i$ para gerar k tópicos e representar a saída pelo conjunto de listas ranqueadas $S_i$.
        ii. Calcular a métrica de concordância $concordância(S_0, S_i)$.
    c. Computar a concordância média para k através das $\tau$ amostras usando $estabilidade(k)$
3. Selecionar um mais valores de k baseados nos maiores valores de estabilidade

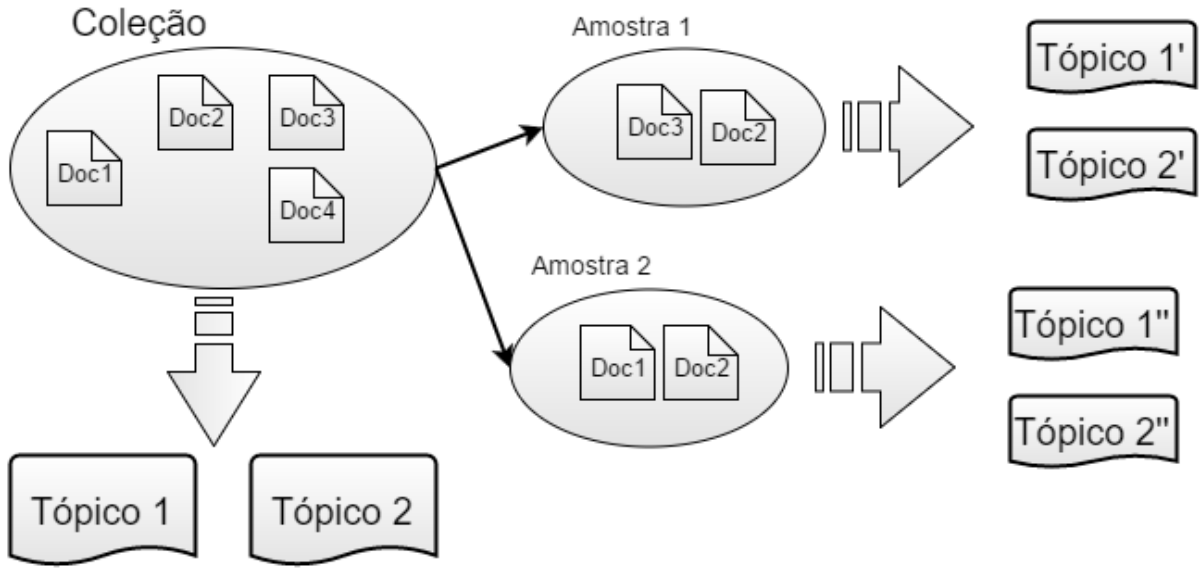



Figura 12: Algoritmo de seleção de grupos com a representação visual da amostragem e extração de tópicos

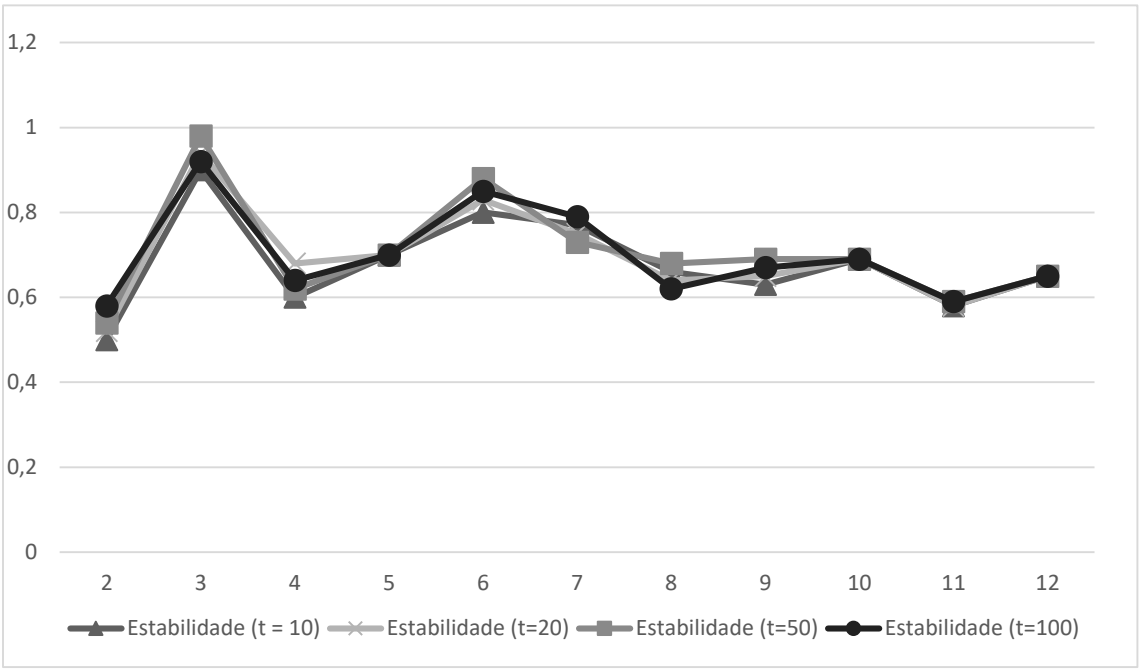



Figura 13: Gráfico dos valores de estabilidade usando t = 10/20/50/100 termos mais relevantes dos tópicos em um corpus de artigos jornalísticos (Adaptado de (GREENE; O'CALLAGHAN; CUNNINGHAM, 2014)).

Quadro 3: Comparativo das principais técnicas para seleção do número de áreas

| **Técnica** | **Características Principais** | **Vantagens** | **Desvantagens** |
|---|---|---|---|
| (LEVINE; DOMANY, 2001) | Cria perturbações nos dados através de amostras. | Amostras normalmente pequenas que se traduzem em uma execução mais rápida | Amostras em conjuntos pequenos podem eliminar certos grupos |
| (BRUNET; TAMAYO; GOLUB, 2004) | Executa o algoritmo várias vezes para verificar perturbações | Permite testar se o máximo global foi alcançado | Execução mais lenta; máximo global pode não ser encontrado |
| (WAAL; BARNARD, 2008) | Executa o algoritmo várias vezes e compara os documentos | Mais precisa pois compara documento a documento no | Mais custosa devido a criação de várias matrizes documento x documento; pode não |

| | | | |
|---|---|---|---|
| | contidos em cada grupo | resultado de cada execução | encontrar um máximo global |
| (NEWMAN et al., 2010) | Mede a coerência entre os termos de um conjunto de tópicos | Métricas avaliam o quão bom estão os principais termos de um tópico | Não considera a robustez quando executado diversas vezes; não é usado para descobrir os parâmetros do algoritmo |
| (GREENE; O'CALLAGHAN; CUNNINGHAM, 2014) | Utiliza as listas de termos já presentes nos tópicos como forma de comparar a estabilidade | Utiliza a própria saída do algoritmo de modelagem como forma de comparar similaridade | Utilizada isoladamente mostra apenas a concordância dos tópicos, necessita o uso aliado as amostras ou múltiplas execuções para um fator de comparação |
| Técnica Escolhida | Utiliza as listas de termos já presentes nos tópicos como forma de comparar a estabilidade | Utiliza a própria saída do algoritmo de modelagem como forma de comparar similaridade; usa amostragem para execução mais rápida e maior controle de precisão | Amostras em conjuntos pequenos podem eliminar certos grupos; |

# 4 Geração de Rótulos

Neste capítulo é apresentado todo o processo de rotulagem dos tópicos que visa representar o conteúdo dos grupos de forma simples e informativa para os usuários. A forma como a modelagem é utilizada atualmente junto com as dificuldades na interpretação dos resultados é abordada na introdução, algumas maneiras de rotular são exibidas na seção de principais técnicas existentes e ao fim é mostrado o processo utilizado neste trabalho na seção de técnicas escolhidas.

## 4.1 Introdução

Após saber o número de áreas existentes na coleção é possível realizar o agrupamento nestas áreas e gerar uma representação por tópico da coleção.

Normalmente, o resultado do agrupamento representa cada grupo com uma distribuição probabilística das palavras mais relevantes para cada um. Um desses resultados pode ser visto na Figura 14.

Podemos utilizar essa lista para que o usuário entenda o assunto e consequentemente a área descrita, o que atualmente é feito na maioria dos trabalhos da literatura (BLEI, 2012; BLEI; NG; JORDAN, 2003; HOFMANN, 1999; LAU et al., 2010). Outra maneira de se descrever a área é utilizar termos que a expressem ou conceitos intimamente relacionados com ela. Esses termos podem ser palavras específicas (bioinformática, agrupamento), pequenas frases com duas a três palavras (redes sociais, mineração de dados) ou até mesmo sentenças (Teoria dos Dois Fatores de Frederick Herzberg). Classificadores humanos frequentemente preferem o uso de frases de duas palavras (CHANG et al., 2009).

O uso da lista resultante do agrupamento muitas vezes é útil para a identificação do assunto. Porém, exige familiaridade com cada área e com o domínio da coleção. Alguém que não domine a temática do domínio pode encontrar dificuldades em interpretar a lista e identificar os conceitos presentes, o assunto principal ou ligar as palavras para formar termos significantes que representem a área em questão. Isso é comum principalmente em áreas de pesquisa onde comumente um tópico é facilmente reconhecido por um grupo familiarizado enquanto dificilmente será reconhecido por outros que não trabalhem especificamente no tema.

Um exemplo pode ser visto na Figura 14, que mostra uma lista resultante com palavras de uma área em ordem de relevância. Para pessoas da área de computação, principalmente os que trabalhem em áreas relacionadas a Sistemas de Informação e Grafos, pode-se inferir que trata-se de documentos relacionados a redes sociais devido a presença dos termos “social” e “redes”. Para quem não vêm da área e portanto não está acostumado a usar essa terminologia pode ser difícil relacionar os termos presentes na lista porque cada um contém mais de um sentido dependendo do contexto. Por exemplo, “social” pode ser usado tanto no contexto de dados quanto no sociológico, econômico ou político. Outros termos como “resultados” e “eficiência” sozinhos não transmitem muita informação útil porque isolados se tornam genéricos.

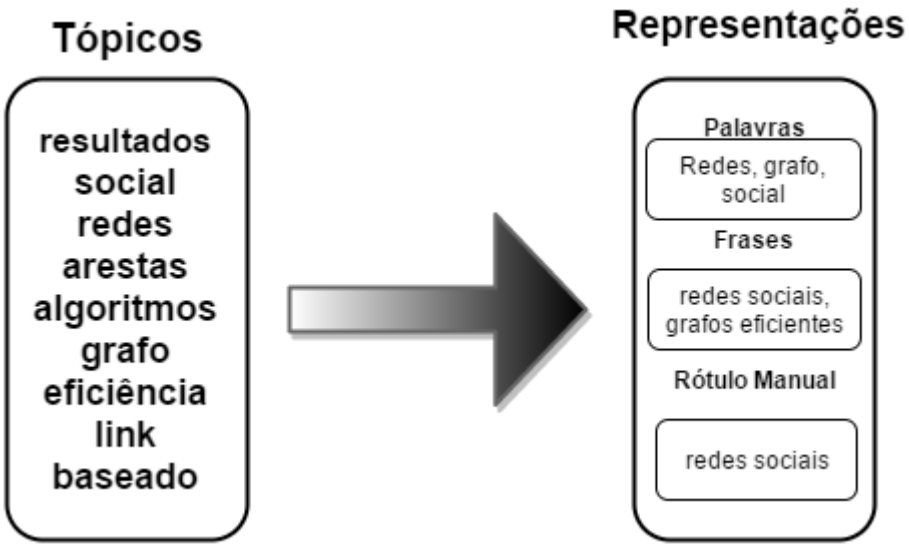



Figura 14: Exemplo de um tópico e suas possíveis representações.

Assim, o desafio da geração de rótulos para as áreas encontradas é de representar cada área de forma automática ao usuário de maneira que se identifique melhor o assunto em questão. Auxiliando na interpretação tanto ajudando o trabalho de quem conhece o domínio quanto facilitando a definição para quem não é familiarizado ao tema.

## 4.2 Principais Técnicas Existentes

A maioria dos trabalhos que utilizam a modelagem de tópicos para agrupamento usam a distribuição de termos de cada tópico extraído como própria representação (BLEI, 2012; HOFMANN, 1999).

Como a lista necessita de uma interpretação que por vezes não é trivial por parte dos usuários, outra opção utilizada é deixar o processo de rotulagem nas mãos de especialistas capazes de gerar rótulos manualmente de acordo com a lista ou com os documentos contidos (CHANG et al., 2009). Essa é uma opção bem confiável visto que os especialistas têm o conhecimento necessário para interpretar e transmitir de forma correta os conteúdos, contudo também não está isenta de interpretações particulares dependendo do *background* do especialista e dificilmente pode ser aplicada em ambientes com grandes volumes de dados e multidisciplinares além de ser custosa e demandar bem mais tempo.

Outras técnicas utilizam abordagens semi-supervisionadas para criação dos rótulos. Nessas técnicas, normalmente o sistema gera rótulos genéricos ou simples e vai refinando o resultado com a ajuda humana.

Como exemplo temos o uso de classificação (LAU et al., 2011; RAMAGE; MANNING; DUMAIS, 2011) que utiliza a modelagem de tópicos como passo não-supervisionado e a própria classificação como passo supervisionado resultando numa abordagem semi-supervisionada. Neste caso, após o agrupamento o sistema gera os rótulos automaticamente baseado em um treinamento da coleção realizado pelos especialistas, que por sua vez devem conhecer as classificações possíveis para o efetivo uso da técnica.

Outra abordagem do tipo seria o aprendizado ativo (DOWNEY et al., 2014), onde o sistema extrai termos para representar a área de forma simples (por exemplo usando algum termo da lista dos mais relevantes) e os especialistas dão um retorno ao sistema de quão bom está aquele rótulo ou melhorando-o e assim se vai modificando o rótulo até que esteja satisfatório.

Portanto, as principais técnicas existentes envolvem o uso de abordagem não-supervisionada, manual ou semi-supervisionada. As não-supervisionadas visam tornar o processo mais rápido as custas da falta do conhecimento especializado. As manuais são as tradicionais, que teoricamente dão melhores resultados utilizando mais recursos tanto pessoal quanto de tempo e as semi-supervisionadas tentam aliviar os problemas das manuais reduzindo o montante de trabalho especialista necessário através da introdução de um passo automático antes do trabalho manual.

## 4.3 Técnicas Escolhidas

Como a proposta visa realizar a identificação das áreas automaticamente, apesar de discutidas as abordagens semi-supervisionadas e os méritos da manual, é necessário se comprometer a utilizar uma abordagem não-supervisionada diminuindo os efeitos que a falta de avaliação por especialistas pode ter nessa geração como a seleção de rótulos mais fáceis de entender e informativos. Outra importante consequência dessa escolha é que dessa forma é possível utilizar a técnica com grandes volumes de dados em menos tempo e que também podemos utilizar dados de diversas fontes e de variados domínios sem necessidade de repetir a mesma carga de trabalho a cada uso.

As principais abordagens totalmente automáticas existentes usam a própria lista para a geração dos termos (por exemplo, os 10 termos mais relevantes) (LAU et al., 2010) ou utilizam alguma estatística dentre todas as palavras presentes na coleção (MEI; SHEN; ZHAI, 2007). Ambas as técnicas não foram satisfatórias para a proposta, devido ao fato de a primeira muitas vezes necessitar de um certo nível de conhecimento na área para interpretação (o que é prejudicial ao propósito de proposta totalmente automática) e da segunda necessitar de tratamentos de texto e processamentos intensivos em toda a coleção, além do fato de desconsiderar termos compostos ou melhor, considerar apenas um tipo de rótulo (palavras ou frases de duas palavras ou três etc.).

Deste modo, foi criada uma nova técnica para este processo que objetiva aliar os benefícios da distribuição probabilística com a análise estatística e a flexibilidade nos tipos de rótulo. A seguir são apresentadas algumas definições básicas de rotulagem e o processo de geração de rótulos em detalhe.

### 4.3.1 Definições

Dada uma coleção de documentos C = $\{d_1, d_2, \dots, d_{|C|}\}$, onde $d_i$ é o documento número *i*, um vocabulário V = $\{w_1, w_2, \dots, w_{|V|}\}$ onde $w_j$ é o termo número j da coleção e um conjunto de tópicos extraídos de C, o objetivo é gerar rótulos compreensíveis para cada tópico (área de pesquisa) que facilitem o entendimento da área.

DEFINIÇÃO 1. Um tópico $\theta$ de C é uma distribuição de probabilidades de termos tal que $\theta = \{ p(w_1|\theta), p(w_2|\theta), \dots, p(w_{|V|}|\theta)\}$ e $\sum_{w \in V} p(w|\theta) = 1$. Assim, termos mais

relevantes para a área teriam maior probabilidade e termos comuns para todas as áreas baixas probabilidades.

DEFINIÇÃO 2. Um rótulo l de $\theta$ é uma palavra ou conjunto de palavras que expressam o conteúdo de $\theta$. Por conseguinte, temos que é possível haver mais de um rótulo possível para cada área já que qualquer palavra usada para exprimir seu conteúdo pode ser utilizada. Isso é visível quando utilizamos sinônimos, embora sejam termos diferentes eles podem ser usados para representar a mesma coisa sem perda de informação.

Finalmente, para selecionarmos rótulos para $\theta$ podemos dividir o processo nas seguintes etapas:

1. Identificar um conjunto de candidatos $L = \{l_1, l_2, \dots, l_n\}$;
2. Calcular S(l, $\theta$), onde S é uma função da relevância do rótulo l para $\theta$;
3. Ordenar os candidatos baseado na função S;
4. Selecionar o(s) rótulo(s) mais relevante(s) para $\theta$ da lista ordenada;

Ao final, os resultados mostram para cada área sua representação por meio de rótulos através desses passos.

### 4.3.2 Processo de geração de rótulos

De acordo com as etapas para criação de um rótulo pode-se dividir o processo em três subtarefas principais: Seleção de candidatos, Ranqueamento, e Seleção de rótulos. A Figura 15 ilustra esse processo.

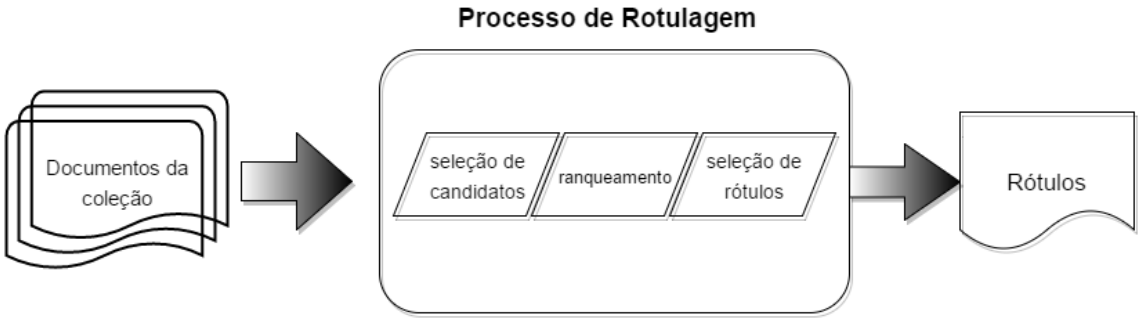



Figura 15: Etapas do processo de Rotulagem

Assim, foram selecionados termos candidatos dos documentos, ordenados por ordem de relevância (segundo uma função para este fim) e selecionados os melhores. Esses três passos são descritos em detalhes a seguir.

#### *4.3.2.1 Seleção de Candidatos*

Para extrair uma lista de candidatos L, antes precisa-se de uma forma de selecionar termos dos documentos de uma área.

Em uma modelagem de tópicos cada documento da coleção tem uma probabilidade associada com cada tópico. A Figura 9 ilustra melhor essa distribuição. Os documentos mais relevantes terão uma maior probabilidade associada com o tópico em questão e uma menor com tópicos não relacionados ou pouco presentes.

Então, para aproveitar a distribuição de probabilidades nos documentos e evitar ruídos é utilizada uma amostra D de documentos de $\theta$. Como os documentos mais relevantes deveriam ser melhores representantes da área, eles serão usados em detrimento da coleção inteira. Isso economiza processamento e torna a abordagem escalável já que independe do número de documentos da coleção.

Utilizando os top-D documentos de $\theta$ pode-se extrair termos que sejam mais relevantes para a área já que estes representam-na melhor. Se a coleção aumentar ou diminuir os candidatos só mudarão se as mudanças afetarem D, novamente tornando mais fácil o uso em ambientes com muitos dados e onde frequentemente há alteração.

Após selecionar o conjunto D já é possível extrair os candidatos para a área. Aqui é utilizado novamente o fato da modelagem de tópicos ser uma abordagem probabilística e vamos utilizar uma amostra W de $\theta$. Como cada tópico pode ser representado por uma distribuição probabilística de termos, são extraídos somente os termos dos documentos que também estão contidos na distribuição. Portanto, para cada termo extraído de D será necessário que ele também esteja contido nos top-W termos de $\theta$ para que seja um candidato válido.

Tanto D como W funcionam como parâmetros para o algoritmo de extração e regulam o quão estrito se quer ser com as associações documento-área e termo-área.

Nas abordagens que utilizam todos os termos da coleção, normalmente o resultado são *stop-words* e termos genéricos em se falando de áreas de pesquisa. Descartando estes, ainda assim sobram muitos termos sem significado para a área. A utilização da amostra D para restringir o número de documentos ajuda a considerar somente os documentos mais relacionados a $\theta$ e quando são extraídos somente os termos contidos em W filtra-se só os candidatos que são relevantes para $\theta$.

Aumentar D faz com que a amostra englobe mais documentos e pode inserir documentos que não tão relevantes para o tópico na mistura, enquanto diminuir pode restringir e deixar os candidatos muito específicos. Da mesma maneira aumentar W consequentemente aumentará o número de rótulos candidatos enquanto diminuir pode deixá-los muito específicos.

Então, a principal parte do algoritmo é a extração de termos e para isso podemos dividir a extração em duas abordagens: **Textual**, que utiliza o corpo de texto como matéria – prima para extração dos termos; e por **Palavras-chave,** que utiliza classificações, termos do autor como forma de descrever o todo.

Uma descrição do algoritmo de seleção de candidatos criado pode ser vista na Figura 17.

**Algoritmo:**

**Entrada: Coleção C e tópico $\theta$ (lista de termos)**

1. **Gerar amostra D de C.**
2. **Gerar amostra W de $\theta$.**
3. **Para cada d $\in$ D:**
    a. **Extrair um conjunto de termos-chave T da amostra D.**
    b. **Para cada t $\in$ T:**
        i. **Se t $\cap$ W:**
            1. **L <= t, onde L é a lista de candidatos**

**Saída: Lista de candidatos L**

Figura 16: Algoritmo de seleção de candidatos.

A seguir são apresentados todos os métodos criados para a seleção nesta pesquisa que foram avaliados para definição de qual seria o melhor na execução da tarefa. Uma comparação entre eles e melhor discussão podem ser encontrados na seção de avaliação.

- **Extração Textual:**

Nesta abordagem será utilizado um algoritmo baseado no *fast keyword extraction algorithm* (BERRY, 2010), que por sua vez é baseado no fato de que os rótulos frequentemente contêm múltiplas palavras mas raramente contém pontuação ou *stopwords.* A entrada para o algoritmo é uma lista de *stopwords* e delimitadores de frases (como pontos e vírgulas). Todas as palavras entre os delimitadores e *stopwords* são consideradas um termo ou rótulo inicial

para uso na seleção de candidatos. As vantagens de se utilizar esse algoritmo como base são sua simplicidade, eficiência, e independência de linguagem, tipo de documento ou domínio. Um exemplo de seu uso pode ser visto no Quadro 4.

Quadro 4: Saída da execução do algoritmo *fast keyword extraction*

| | |
|---|---|
| Texto Original | Temporal Text Mining (TTM) is concerned with discovering temporal patterns in text information collected over time. Since most test information bears some time stamps, TTM has many applications in multiple domains, such as summarizing events in news articles and revealing research trends in scientific literature. |
| Saída do Algoritmo | “time stamps”, “summarizing events”, “discovering temporal patterns”, “news articles”, “concerned”, “test information bears”, “applications”, “test information collected”, “temporal text mining”, “TTM”, “multiple domains”, “scientific literature”, “revealing research trends”, “time” |

Após essa extração os termos são utilizados normalmente no algoritmo de seleção de candidatos. Para efeitos de avaliação foram criadas variantes dessa extração textual, visando eleger a melhor extração possível. Assim, para facilitar o entendimento, essa extração será referida como **ET1** (Extração textual 1) e suas variantes como:

**ET2**: Nessa abordagem, quando o termo extraído possui mais de duas palavras, estas são divididas em bigramas e continua-se o algoritmo incluindo estes bigramas na lista de candidatos se contiverem algum termo de W. Essa variação foi criada baseada na preferência dos indexadores humanos por termos de duas palavras (CHANG et al., 2009).

**ET3:** Da mesma forma, divididem-se em bigramas quando houver mais de duas palavras, mas se houver mais que três, estas serão divididas em trigramas como uma forma de favorecer frases curtas no conjunto.

- **Extração por palavras-chave:**

Ao contrário da textual, na extração por palavras-chave são extraídas palavras-chave descritas pelos autores dos documentos assim como descritores e classificações quando presentes. Da mesma forma que na extração textual aqui são utilizados **EP1** como uma sigla

para a extração que utiliza as palavras-chave presentes nos documentos como termos para a seleção de candidatos e:

**EP2:** Em adição as palavras-chave, são adicionados os descritores como candidatos (por exemplo, taxonomias presentes).

**EP3:** Semelhante a **ET2**, são gerados bigramas e trigramas para cada termo pelos mesmos motivos.

É importante frisar que nem todas as coleções possuirão rótulos providos pelos autores muito menos descritores. Inicialmente pode-se pensar que as palavras-chave são a melhor forma de descrever a área já que são relevantes aos documentos. Infelizmente, como o objetivo é rotular uma área de pesquisa, as palavras-chave muitas vezes são específicas demais para essa tarefa, já que se aplicam ao documento, e os descritores muito genéricos e amplos, reduzindo sua capacidade de descrever os conceitos da área.

#### *4.3.2.2 Ranqueamento*

Após a produção de termos candidatos para uso como rótulos das áreas, o passo seguinte é ordená-los de acordo com a relevância de cada um para o grupo. Da mesma maneira que a seleção de candidatos, aqui foram utilizados alguns métodos de ranqueamento existentes e criados alguns novos. Uma comparação entre os resultados e melhor discussão pode ser encontrada na seção de avaliação.

Das técnicas testadas foram utilizadas duas já existentes: A **Frequência de termos** (*tf*) e a **Relação Grau/Frequência** (*deg/tf*) e conforme foi-se testando empiricamente os resultados, adaptou-se conceitos de grau dos termos e tamanho para criar outras duas: O **Grau modificado de rótulo** e o **Grau modificado estendido**. A motivação e explicação do uso destas técnicas é detalhada nas próximas subseções.

- **Frequência de Termos:**

A frequência de termos (normalmente representada pela sigla tf, do inglês *term frequency*) é uma forma tradicional e muito utilizada para atribuir uma pontuação aos termos dependendo de sua relevância em relação ao corpus. Se baseia na suposição de que o peso de um termo que ocorre em um documento é diretamente proporcional à sua frequência (LUHN, 1957).

Assim, para calcular a frequência, basicamente se conta a quantidade de vezes que um termo aparece em um documento. Para o uso na geração de rótulos, toma-se todos os termos selecionados como candidatos e calcula-se a frequência de cada um utilizando os documentos mais relevantes da amostra como fonte textual.

Normalmente a frequência, como pode-se esperar, dá pontuações altas para *stop words* e termos que não são muito descritivos (como verbos comuns). Por isso, costuma-se usar o **inverso da frequência nos documentos** (*idf*) (SPARCK JONES, 1972) ou a relação ***tf-idf*** (RAJARAMAN; ULLMAN, 2011). O primeiro dá mais peso a termos que ocorrem mais raramente, enquanto o segundo dá valor ao número de ocorrências, no entanto, esse valor é equilibrado pela frequência da palavra no corpus.

Como são removidas as *stop words* na fase de seleção dos candidatos, não há o primeiro ponto negativo do uso da frequência. Adicionalmente, são selecionados apenas os candidatos que estão presentes na lista dos termos mais relevantes W, de acordo com o algoritmo de modelagem de tópicos, assim removendo um segundo conjunto de termos irrelevantes. Então, a princípio o uso da frequência neste caso estaria isento de seus pontos negativos. Como o objetivo é de que os rótulos representem descrições da área, não seria possível o uso do inverso da frequência nos documentos por sua especificidade e devido ao tamanho da amostra dos documentos relevantes usada para a seleção e ranqueamento, não há muito ganho na utilização do *tf-idf*.

Além da filtragem de termos que é feita na seleção de candidatos, a frequência é capaz de remover agora os termos mais longos e infrequentes e dar mais ênfase em termos mais curtos e frequentes. Esses termos podem ser mais gerais e não muito descritivos, porém o ranqueamento é feito utilizando a amostra de documentos mais relevantes para a área.

- **Relação grau/frequência:**

Se por um lado a frequência favorece termos mais curtos, que tendem a aparecer mais, a relação grau/frequência (*deg/tf*) pode favorecer termos mais longos combinando os conceitos de grau (*deg*) e frequência (*tf*).

O grau (normalmente representado pela sigla *deg,* do inglês *degree*) de uma palavra é definido como a quantidade de vezes em que ela aparece isolada (neste caso, palavra = termo) somado com a quantidade de vezes em que ela aparece incluída em um termo (para termos

com mais de uma palavra). Para termos, o grau é calculado como a soma dos graus de suas palavras (BERRY, 2010).

Como o grau tende a favorecer as palavras que ocorrem em candidatos mais extensos (já que para termos é a soma dos graus das palavras) e a frequência termos com alta ocorrência, que costumam ser mais curtos, a relação grau/frequência favoreceria termos que ocorrem predominantemente em candidatos mais longos.

Essa métrica é uma forma de beneficiar termos frequentes tanto isoladamente quanto quando aparecem como parte de um termo maior.

- **Grau modificado de rótulo:**

Recordando as definições de grau, temos que o grau de uma palavra é a sua frequência como termo somada a frequência em que aparece dentro de termos compostas. O grau de um termo é simplesmente a soma dos graus de suas palavras constituintes. Estendendo esses conceitos e adaptando-o para o uso com os rótulos é definido aqui o conceito de **grau do rótulo** (*ldeg*) que basicamente é a frequência com que um rótulo candidato aparece na lista de candidatos somada a frequência com que esse rótulo aparece incluído em outros candidatos. A diferença entre o grau do rótulo e do termo seria que para os rótulos se considera o termo como um todo e não suas palavras. Assim, o grau de um termo seria a soma do grau de suas palavras enquanto o grau do rótulo seria a quantidade de ocorrências de um termo isoladas ou como parte de outro, como se o próprio termo fosse uma palavra.

Essa definição beneficia principalmente unigramas (termos compostos por uma palavra) pois tendem a aparecer mais frequentemente. Por exemplo, um termo como “dados” pode aparecer em “mineração de dados”, “visualização de dados”, “análise de dados”, mesmo que cada um represente um conceito diferente.

Uma solução para esse problema seria balancear as pontuações tanto de unigramas quanto de n-gramas (termos compostos por n palavras), de preferência fazendo um ajuste tal que se diminuísse o peso dos unigramas e se aumentasse o peso dos n-gramas, principalmente dos que aparecem isolados e não como parte de outro termo.

Então, baseado na extensão da definição de graus para rótulos e levando em conta o balanceamento entre termos simples e compostos, foi criada a métrica de **grau modificado do rótulo** (de agora em diante representada pela sigla *mdeg* para melhor compreensão

quando comparada com outras métricas existentes). Ela pode ser definida para um rótulo l como:

$$mdeg(l) = ldeg(l) + tf(l)$$

mdeg(l) = Número de ocorrências como parte de um termo composto + 2*tf(l)

Equação 7: Grau modificado do rótulo

Ou seja, o grau modificado de um rótulo é a soma do grau do rótulo e de sua frequência. Para cada ocorrência do termo como parte de outro atribui-se um ponto. Se os dois termos comparados são iguais atribui-se dois pontos. Agora, comparando os termos "dados" e "mineração de dados" com um termo "mineração de dados" daria uma pontuação de um ponto para "dados" (*match* parcial) e dois para "mineração de dados" (*match* perfeito).

Essa métrica poderia casar bem com as técnicas de seleção de candidatos ET2 e ET3 por exemplo, que criam bigramas e trigramas a partir dos termos mais extensos. Por fim, essa métrica visa dar mais peso para termos que aparecem isolados sem remover completamente o peso de unigramas e termos muito frequentes.

- **Grau modificado estendido:**

Outra forma de balancear o grau do rótulo seria diminuindo os pesos tanto de unigramas quanto de termos longos, favorecendo termos compostos por pequenas frases, normalmente de duas a três palavras. Essa pontuação maior para frases curtas poderia pontuar melhor rótulos mais significativos para as pessoas, visto que de acordo com o trabalho de (CHANG et al., 2009), ao indexar documentos manualmente, os indexadores preferem frases curtas ao uso de palavras e de frases longas. Claramente, algumas áreas podem ser descritas facilmente por uma palavra, enquanto outras necessitam de termos maiores. Então, o objetivo aqui não é eliminar a relevância de palavras e termos longos e sim balanceá-los dando um peso um pouco maior a termos compostos por poucas palavras já que são preferidos quando usados por humanos.

Considerando-se isso foi criada a métrica de **grau modificado estendido** (será utilizada a sigla *medeg* de agora em diante para facilitar a compreensão e comparação) que utiliza o balanceamento entre termos simples e compostos beneficiando frases curtas e penalizando frases muito longas.

Essa nova métrica pode ser definida formalmente dado um rótulo l por:

$$medeg(l) = (tf(l) + ldeg(l)) * (1 + \log(wc(l)))$$

Equação 8: Grau modificado estendido

Onde wc(l) é o número de palavras presente em l. O primeiro termo à direita da equação corresponde ao balanço que beneficia termos únicos com mais pontos, somando o grau do rótulo e sua frequência. O segundo termo equilibra os pesos dependendo do tamanho do rótulo (em palavras). O uso da função logarítmica neste caso visa equilibrar as pontuações dos rótulos mais longos, evitando valores exorbitantes, enquanto dá um peso um pouco menor para unigramas. Levando em conta que o log(2) = 1, tem-se que a adição da curva logarítmica tende a suavizar as pontuações tanto de termos com wc = 1, que em teoria deveriam ser mais frequentes, quanto termos com wc alto, que não teriam pontuações muito altas devido a curva suave, contrária à exponencial.

Finalmente, no Quadro 5 é exibida uma comparação entre as formas de ranqueamento usadas e testadas junto com as respectivas pontuações em um caso exemplo.

Quadro 5: Comparação das pontuações das métricas apresentadas.

| **Rótulos Candidatos** | **Pontuações para o candidato: “social networks”** | | | |
|---|---|---|---|---|
| “social networks”, “social networks systems”, “social networks applications”, “learning algorithms”, “social”, “classifier” | tf | mdeg | medeg | deg/tf |
| | 1 | 4 | 5.20 | 3 |

#### *4.3.2.3 Seleção de Rótulos*

Depois de realizar o ranqueamento dos rótulos, o último passo é selecionar para cada área um deles e exibir como representante do assunto. Como a lista de rótulos já está ordenada, para ter apenas um rótulo para a área basta selecionar o primeiro da lista. O problema na seleção de rótulos surge quando se usa múltiplos rótulos (mais de um rótulo descrevendo a área). Quando utiliza-se vários rótulos, cada um dos rótulos escolhidos necessita ao menos representar uma visão distinta dos conceitos englobados pela área, ao invés de serem sinônimos entre si.

Para solucionar as peculiaridades do uso de múltiplos rótulos, são definidos antes dois tipos de seleção que podem ser utilizados em conjunto ou separadas para este caso: Seleções inter-tópico e intra-tópico. Essas duas formas de seleção são inspiradas nos conceitos de seleção inter e intra *cluster* (MANNING et al., 2007), obviamente adaptando-se ao cenário de modelagem de tópicos que apresenta um outro paradigma de grupos. Essas duas definições são detalhadas nas próximas subseções

- **Seleção inter-tópico:**

A seleção inter-tópico é usada quando existem interseções nos rótulos de áreas diferentes, ou seja, o mesmo rótulo aparece em dois tópicos.

Como exemplo, suponha que existam dois tópicos extraídos pela modelagem chamados $\theta_1$ e $\theta_2$ com dois conjuntos de rótulos L1 = $\{l_1, l_2, l_4\}$ e L2 = $\{l_3, l_1, l_5\}$ respectivamente dos quais deseja-se selecionar dois rótulos para cada tópico. Neste caso, o rótulo $l_1$ se encontra na primeira posição do conjunto final de rótulos L1 e também na segunda posição do conjunto L2 (rótulos de L1 e L2 ranqueados na ordem de leitura). Quando selecionados haveria o mesmo rótulo em ambos os tópicos, porém, devido ao ranqueamento, sabe-se que $l_1$ é mais relevante para $\theta_1$ do que para $\theta_2$(devido a posição na lista ordenada). Para evitar o problema da interseção na representação das áreas, diferenciando-as o máximo possível, a abordagem utilizada neste caso seria a de atribuir o rótulo ao tópico mais relevante correspondente selecionando um outro na ordem que distinga melhor as áreas. No caso de exemplo, atribuir-se-ia $l_1$ à $\theta_1$ devido a sua maior relevância ao tópico, selecionando $l_1$ e $l_2$ como rótulos para $\theta_1$. Para $\theta_2$, seriam selecionados $l_3$ e $l_5$, devido ao fato da relevância de $l_1$ser menor para este tópico e escolhendo então o rótulo seguinte respeitando-se a sequência do ranqueamento ($l_5$, neste caso). Quando há mais de um rótulo idêntico entre áreas, pode-

se repetir o processo até que se ache rótulos suficientemente diferentes entre as áreas ou até que se termine a lista.

O caso de um mesmo rótulo aparecer em tópicos diferentes pode ocorrer pois cada documento é modelado como uma mistura de tópicos, então é possível que um rótulo de uma área apareça como sugestão de outra na seleção, principalmente se as áreas não forem suficientemente diferenciadas.

Apesar da definição criada, este tipo de seleção não será utilizado neste trabalho, pois assume-se que a modelagem de tópicos e a seleção do número de áreas ideais é capaz de diferenciar as áreas de maneira que não haja intercessões relevantes entre elas. Os filtros usados na seleção de candidatos também restringem os rótulos possíveis aos termos mais relevantes para cada determinado tópico, diminuindo ainda mais a possibilidade da ocorrência de um mesmo rótulo como relevante para duas áreas distintas. Outra desvantagem do uso de seleção inter-tópico é a de que a seleção de rótulos para um tópico depende dos rótulos de todos os outros tópicos. Essa característica torna este tipo de seleção mais trabalhoso para coleções onde estão presentes um grande número de áreas ou para quando se usa coleções dinâmicas, onde pode-se adicionar ou remover documentos da coleção após o processamento (neste caso seria necessário realizar todas as operações novamente para a coleção inteira).

- **Seleção intra-tópico:**

Como pode-se imaginar, a seleção intra-tópico é realizada para selecionar a melhor sequência de rótulos dentro de um mesmo tópico visando facilitar ao máximo seu entendimento.

Suponha que sejam selecionados dois rótulos para um determinado tópico e por ordem estes sejam “mineração de dados” e “mineração” respectivamente. Claramente, um segundo rótulo “mineração” é redundante para uma área onde o assunto é mineração de dados, já que são quase que sinônimos. Outros rótulos, como “algoritmos” ou “aprendizado” podem oferecer outras perspectivas sobre os assuntos específicos da área. Esse é o objetivo deste tipo de seleção, eliminar sinônimos e termos redundantes favorecendo rótulos adicionais mais esclarecedores do assunto tratado.

Para realizar esta seleção, este trabalho adota uma abordagem simples que visa a eliminar termos textualmente semelhantes removendo rótulos que estejam incluídos em

outros de maior relevância. Como exemplo, dado um conjunto de rótulos L1 = $\{l_1, l_2, l_3\}$ associado a um tópico $\theta$, se $l_1$ contém $l_2$ como parte de si, substitui-se $l_2$ (pois possui menor relevância para $\theta$) por $l_3$ (o próximo rótulo de acordo com a ordem). Então, se aqui houvessem os rótulos “mineração de dados” e “mineração” o segundo seria substituído pelo próximo da lista por ter menos relevância que o primeiro de acordo com a pontuação e por estar contido também no primeiro.

Após feitas as seleções toda a geração de rótulo está finalizada com cada área associada a seu(s) rótulo(s).

Quadro 6: Comparativo das técnicas para geração de rótulos

| **Técnica** | **Características Principais** | **Vantagens** | **Desvantagens** |
|---|---|---|---|
| Utilizar lista truncada de termos do tópico | Utiliza a própria lista ordenada por relevância do tópico | Nenhum processamento adicional; lista já ordenada por relevância ao tópico | Difícil interpretação; termos associados ao tópico nem sempre são semanticamente associados |
| Manual | Utiliza interpretação por especialistas | Maior precisão; uso do conhecimento no domínio | Mais custoso tanto em recursos humanos quanto em tempo de execução |
| Semi-supervisionada por classificação | Utiliza uma classificação para associar cada tópico ao rótulo correspondente | Acerto maior pois os tópicos já são pré-determinados | Exige participação de especialistas para definição e treino dos classificadores |
| Semi-supervisionada por aprendizado ativo | Cria rótulos primitivos e refina baseado no retorno dado por especialistas | Rótulos em evolução, sempre melhoram quanto maior o retorno dos especialistas | Exige conhecimento humano especializado; necessita de um certo tempo para que os |

| | | | |
|---|---|---|---|
| | | | rótulos se tornem adequados |
| Técnica Escolhida | Utiliza amostragem nos termos e documentos para considerar somente os mais relevantes; usa palavras-chave para representar os grupos | Usa as palavras-chave combinadas à distribuição de termos e tópicos natural da modelagem; escalável para coleções grandes | Depende dos termos presentes no texto; exclui termos que não estejam dentro da amostra |

# 5 Análise Temporal

Até então, foram mostradas as etapas da proposta que pretendiam agrupar documentos por área a partir de uma coleção, selecionar o número de áreas presentes na mesma coleção e rotular cada área de pesquisa encontrada respectivamente.

Nesta seção, serão apresentadas técnicas que extrapolam o uso do processo descrito na proposta para a realização da identificação das áreas temporalmente. Seja por ano, quinquênio ou até mesmo meses, pode-se realizar uma análise das áreas com maior crescimento, declínio, divisão ou fusão ao longo do tempo.

## 5.1 Introdução

O uso da proposta aliado ao fator tempo presente naturalmente em coleções científicas pode ser visto como uma aplicação de mineração temporal no corpus. A mineração temporal de textos (TTM) é uma área cujo objetivo é descobrir a estrutura latente e padrões temporais em coleções de texto. Estas características são importantes em coleções em que os tópicos de interesse mudam frequentemente com o passar do tempo, como frequentemente ocorre na ciência e tecnologia. Além disso, a mineração temporal de textos é útil em ferramentas de sumarização e descoberta de tendências.

A utilização da proposta com a mineração temporal permite que sejam encontrados padrões entre áreas no mesmo intervalo de tempo e entre dois intervalos de tempos distintos. Assim, se pode inferir como um tópico ou área influenciou toda a pesquisa posterior ou os documentos que deram origem a um campo de estudo bem estabelecido atualmente dada uma coleção que abarque todo o período de vida das áreas. Um exemplo visual pode ser visto na Figura 18.

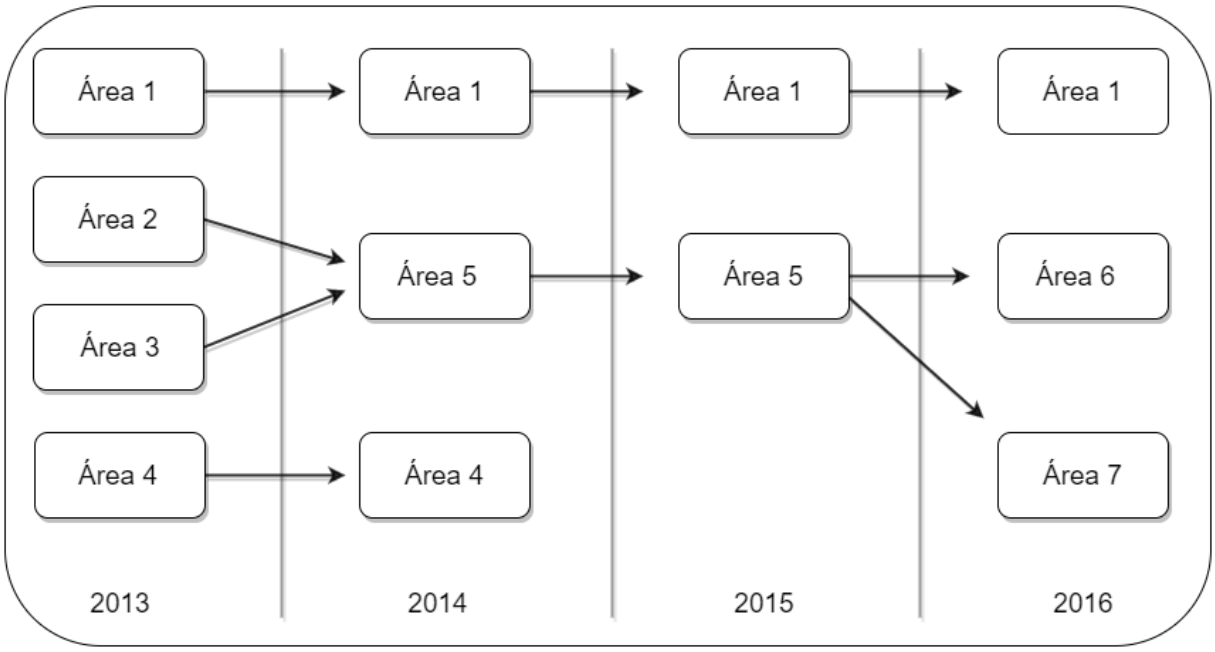



Figura 17: Exemplo de análise temporal de áreas

A seguir são apresentadas as técnicas existentes e exemplos de análise temporal em coleções textuais e a abordagem utilizada neste trabalho para alcançar este fim tendo em vista as características do modelo adotado.

## 5.2 Técnicas Existentes

Basicamente, como a proposta utiliza a modelagem de tópicos como forma de agrupar documentos por temas, as técnicas de mineração temporal que podem ser utilizadas se limitam às que trabalham neste contexto.

As abordagens presentes atualmente na literatura dividem-se nas que utilizam novos algoritmos de modelagem de tópicos para considerar a dimensão tempo (KAUER, 2013; MEI; ZHAI, 2005) e nas que utilizam processamentos posteriores na coleção ou que ignoram a modelagem realizada trabalhando unicamente nos documentos fonte (KUROSAWA; TAKAMA, 2011; PRIYA; KUMARAVEL, 2013), muitas vezes através de visualização ou análise de citações.

### 5.2.1 Algoritmos de modelagem Dinâmicos

A maioria dos trabalhos em modelagem de tópicos não leva em consideração as estruturas temporais dos documentos (KAUER, 2013; MEI et al., 2006).

Os principais algoritmos para a obtenção de modelos dinâmicos incluem (MEI; ZHAI, 2005) que realiza uma modelagem na coleção completa e um posterior agrupamento para

cada período de tempo, estabelecendo ao fim uma relação entre o agrupamento ao longo de todo o tempo e o agrupamento realizado nas fatias de tempo. Esse algoritmo é melhor utilizado quando os temas são fortes ao longo de todo o tempo utilizado na análise. Quando estão presentes temas que foram salientes durante um curto período de tempo este tende a desaparecer quando comparado ao período total da coleção.

Outro algoritmo dinâmico mais conhecido é o presente em (BLEI; LAFFERTY, 2006) que realiza um agrupamento em cada período de tempo (por anos, por exemplo) e constrói tópicos de um determinado período a partir do que estava presente no período anterior. Neste caso, temas fortes no passado influenciam o agrupamento realizado no presente.

Vale ressaltar que a maioria dos algoritmos dinâmicos sofrem do problema da seleção de áreas da mesma maneira que os estáticos. O usuário no caso deve ter uma ideia da quantidade de temas presentes na coleção antes de utilizá-los e além disso a quantidade de temas presentes em cada período de tempo.

### 5.2.2 Algoritmos independentes

Muitas vezes não é possível ou ideal realizar uma nova modelagem de tópicos em coleções que já possuem uma, seja por motivo de tempo, custo ou tamanho da base. Aqui então entram técnicas mais modulares que podem ser utilizadas em conjunto com qualquer algoritmo de modelagem de tópicos, porém sem se aproveitar da modelagem realizada.

O trabalho de (PRIYA, M. B., KUMARAVEL, 2013) utiliza citações entre documentos na internet para tal fim, agrupando-os por período e depois extrapolando as citações entre documentos para citações entre grupos. (KONTOSTATHIS et al., 2004) realiza uma pesquisa de técnicas para detectar assuntos emergentes em coleções textuais porém nenhuma delas é capaz de realizar uma análise evolucionária nos moldes do que é visto na Figura 18.

Outra forma de realizar  a análise temporal é por meio de visualização como é usado em (CHEN, 2006b; KUROSAWA; TAKAMA, 2011). Desta forma não é necessário um processamento automático para inferir relações temporais, exibindo os padrões que aparecem em cada instante de tempo e delegando a interpretação aos usuários ou especialistas.

Por fim, no escopo científico, vale destacar o trabalho de (JANSSENS; GLÄNZEL; MOOR, 2007) que utiliza a análise de citações acadêmicas na área de bioinformática para a descoberta

de padrões temporais. A restrição nesse caso é que a coleção presumivelmente deve englobar os que citam e os citados para uma análise completa.

## 5.3 Técnicas Escolhidas

Como mostrado na seção anterior, existem duas abordagens principais para realizar uma análise temporal no contexto da modelagem de tópicos. Uma se utiliza de novos algoritmos customizados para cenários dinâmicos e a outra de técnicas independentes do algoritmo de modelagem utilizado.

Tendo em vista a flexibilidade da proposta apresentada neste trabalho e sua modularidade, optou-se pela utilização de uma abordagem que não esteja fortemente acoplada ao modelo de tópicos utilizado. Os algoritmos dinâmicos são relativamente recentes e ainda não tão consagrados e avaliadas como os tradicionais. Além disso exigem um conhecimento mais amplo por parte de quem usa, pois este deve ter uma noção da quantidade de áreas por período analisado para ajustar os parâmetros dos algoritmos. Como o objetivo é uma identificação automática, o ideal é evitar ao máximo a adição de elementos que dependam do conhecimento humano.

Entre as técnicas que são independentes, muitas usam a informação de citação ou as ligações entre os documentos. Novamente, como o trabalho usa áreas de pesquisa como matéria-prima, nem sempre é possível conter uma rede completa de citações dentro de uma coleção. Por exemplo, se o objetivo é analisar a evolução das áreas de pesquisa dentro de uma conferência ou de uma universidade, não é plausível acreditar que essas áreas englobem todo o universo de temas e documentos presentes nas mesmas áreas a nível mundial.

Então, ao invés de utilizar redes de citações e outras estruturas externas para incluir o elemento dinâmico na coleção e nas áreas encontradas, optou-se por criar uma técnica que visa realizar essa mesma tarefa aproveitando-se da estrutura da modelagem de tópicos e da seleção de áreas já usadas nas etapas da proposta.

Nas seções seguintes são apresentadas as definições e a técnica criada respectivamente, ambos para o fim da análise temporal do corpus. Primeiro são definidos os conceitos necessários para a adição do elemento temporal ao processamento de identificação de áreas de pesquisa. Após as definições, é apresentada uma técnica que utiliza os próprios grupos e probabilidades oriundos da modelagem de tópicos como forma de criar uma visão temporal da coleção.

### 5.3.1 Definições

Dada uma coleção de documentos C = $\{d_1, d_2, \dots, d_{|C|}\}$, onde $d_i$ é o documento número *i*, um vocabulário V = $\{w_1, w_2, \dots, w_{|V|}\}$ onde $w_j$ é o termo número j da coleção e um conjunto de tópicos extraídos de C, o objetivo aqui é realizar a extração dos tópicos ao longo do tempo, permitindo uma análise evolucionário dos temas de pesquisa.

DEFINIÇÃO 1. Um **tópico** $\theta$ de C é uma distribuição de probabilidades de termos tal que $\theta = \{ p(w_1|\theta), p(w_2|\theta), \dots, p(w_{|V|}|\theta)\}$ e $\sum_{w \in V} p(w|\theta) = 1$. Assim, termos mais relevantes para a área teriam maior probabilidade e termos comuns para todas as áreas baixas probabilidades.

DEFINIÇÃO 2. Um **intervalo temático** é um intervalo temporal no qual um tópico aparece emergindo de C. O intervalo pode ser representado como $(\theta, t_i, t_f)$, onde $\theta$ é o tópico, $t_i$ o início do intervalo de tempo e $t_f$ o fim do intervalo de duração.

DEFINIÇÃO 3. Uma **transição evolucionária** é uma relação de similaridade entre dois intervalos temáticos. Dados $(\theta_1, t_0, t_1)$ e $(\theta_2, t_2, t_3)$, se $t_1 \leq t_2$ (ou seja, o primeiro tópico começa antes do segundo iniciar) e esses tópicos possuem uma alta similaridade, pode-se dizer que $\theta_2$ evolui de $\theta_1$.

DEFINIÇÃO 4. Um **grafo de evolução dos tópicos** é um grafo direcionado com pesos G = (N, E) onde cada vértice N é um intervalo temático (um tópico contido num intervalo de tempo) e cada aresta E é uma transição evolucionária (ligação entre tópicos). As arestas possuem peso para indicar a força da transição ou a distância evolucionária. Pesos maiores indicam uma maior proximidade entre as áreas. Áreas que se ramificam por exemplo, tendem a ter pesos iguais entre seus ramos

Assim, o problema da análise temporal resume-se em encontrar e criar um grafo de evolução dos tópicos a partir de uma coleção de documentos. O processo pode ser dividido nas seguintes etapas:

1. Dividir a coleção em intervalos de tempo Δ (dependendo da natureza da coleção e da análise esse intervalo pode ser de anos, meses, décadas...).
2. Extrair os tópicos de cada parte da coleção dividida entre os intervalos, constituindo os intervalos temáticos.
3. Encontrar as transições evolucionárias entre os tópicos através de uma função de similaridade.

4. Construir o grafo de evolução dos tópicos a partir dos intervalos e transições.

Um exemplo do grafo final pode ser visto na Figura 19. As áreas de pesquisa seriam os tópicos extraídos, t-1, t e t+1 os intervalos de tempo e arestas indicam a evolução entre as áreas. Daqui se entende que a área i por exemplo se mantém relevante ao longo do tempo, enquanto outras se limitam a intervalos específicos.

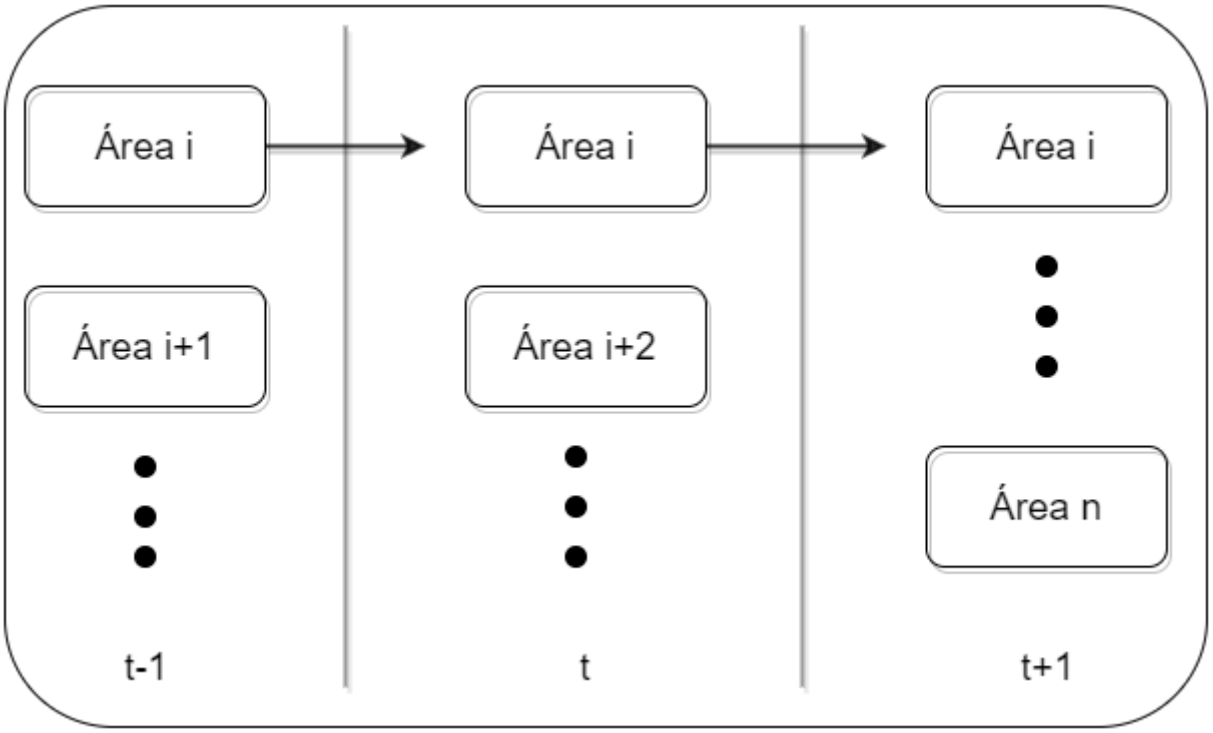



Figura 18: Grafo de evolução de tópicos. No cenário científico, as áreas de pesquisa são os tópicos

Na próxima seção é apresentado o processo utilizado neste trabalho para a obtenção desse grafo para a análise temporal das áreas de pesquisa. São apresentadas as técnicas utilizadas para a realização das etapas descritas anteriormente, assim como a forma de atribuir similaridade e de dar pesos as arestas do grafo.

### 5.3.2 Grafo de evolução dos tópicos

Para permitir a análise temporal das áreas de pesquisa, é criado um grafo sobre a coleção que mostra a sua evolução temporal por meio dos tópicos extraídos. Para realizar isso, seguem-se as etapas definidas na seção anterior.

Primeiramente deve-se dividir a coleção em intervalos temporais, por exemplo, uma coleção de artigos, patentes ou de livros normalmente é dividida por ano. Nada impede que

a coleção seja dividida em intervalos menores como meses ou em maiores como décadas, isso dependerá do tipo de análise e da natureza do corpus.

Tendo em vista que documentos da ciência e tecnologia são formais e comumente possuem uma data associada, essa primeira tarefa é simples. Após a divisão, a coleção se torna $C = \{c_1, c_2, \dots, c_{|C|}\}$, onde cada $c_i$ é um subconjunto de C contendo os documentos pertencentes a um intervalo t (um ano, por exemplo).

Com a coleção dividida em fatias de tempo, segue-se a extração dos tópicos para cada subconjunto $c_i$. Aqui, simplesmente será executada proposta e suas etapas de agrupamento, seleção de áreas e rotulagem respectivamente. Tendo executados esses passos obtém-se os intervalos temáticos, que são os temas de pesquisa presentes em cada intervalo de tempo.

Depois de obter as áreas presentes em cada intervalo de tempo, deve-se encontrar as transições evolucionárias entre elas. Alguma forma de obter uma similaridade entre elas é então necessária.

Neste passo o que foi feito foi uma execução do agrupamento e seleção de áreas unindo-se dois subconjuntos $c_i$ sequenciais no tempo. Se o intervalo de tempo utilizado é anual e foram extraídos os tópicos dos documentos de 2015 ($c_1$) e de 2016 ($c_2$), agora extraem-se também tópicos dos documentos de 2015-2016 (ou seja, de $c_1 + c_2$), por exemplo.

Como cada tópico possui um conjunto de documentos associados, a similaridade entre tópicos é tomada como a quantidade de documentos em comum que estes possuem quando agrupados juntos. Se um tópico $\theta_1$ está presente em 2015 e um outro tópico $\theta_2$ em 2016, a similaridade entre eles será a interseção de seus documentos nos tópicos presentes 2015-2016. Uma ilustração melhor desse conceito pode ser vista na Figura 20.

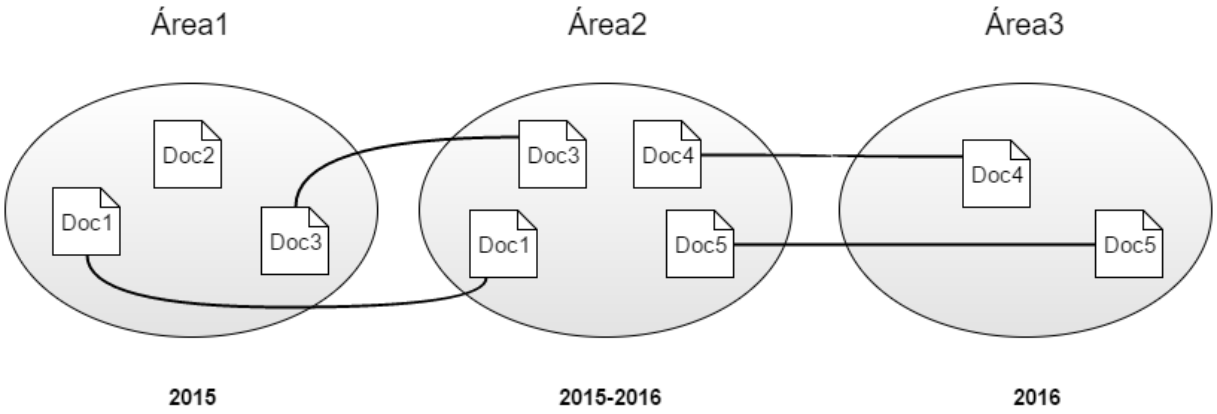



Figura 19: Exemplo de obtenção da transição entre áreas num intervalo anual.

Assim, se uma área se mantém ao longo do tempo, os tópicos de períodos sequenciais vão conter todos os documentos do período anterior e do período seguinte (no exemplo, 2015-2016 conterá os documentos dos tópicos em 2015 e em 2016). Caso contrário, pode ter havido uma diminuição na força da área naquele período (se houve diminuição no total de documentos nos períodos sequenciais), um aumento (se o total aumentou) ou uma ramificação (um tópico de interesse se dividiu em dois ou mais novos interesses). É possível até mesmo que em um determinado intervalo de tempo uma área não esteja presente (por exemplo, quando se perde o interesse na pesquisa) ou que apareçam áreas novas (por exemplo, quando há alguma descoberta nova). Todos esses casos dependerão da proporção de documentos que passa entre as áreas presentes nos dois intervalos através da interseção entre eles.

Essa relação obtida entre as áreas é vista como a transição evolucionária entre elas. Apesar de possuir documentos diferentes, quando agrupados em conjunto as duas áreas encaixam-se no mesmo grupo o que significa que a área no período de tempo mais recente é uma evolução ou continuação da mesma área do período anterior.

A Figura 21 mostra um exemplo das transições obtidas com um intervalo de tempo anual no período de dois anos.

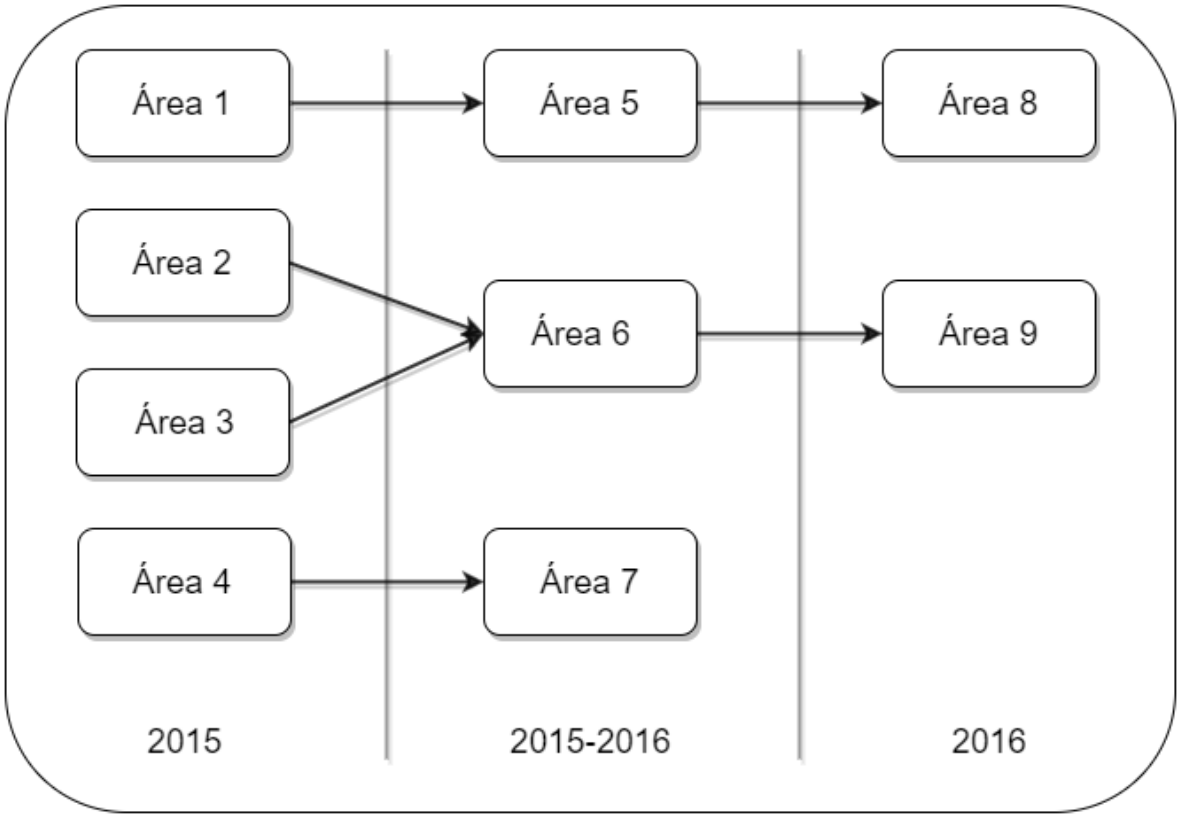



Figura 20: Exemplo de transições obtidas em várias áreas no período de dois anos

A partir daí a obtenção do grafo de evolução dos tópicos é trivial, já que as arestas e vértices já foram obtidos. Os agrupamentos sequenciais (2015-2016) são temporários e serão removidos e o que fica é a transição entre as áreas iniciais e finais para cada período de acordo com a fatia de tempo utilizada.

Uma observação importante acerca do grafo é de que suas arestas têm peso, esse peso é dado pela força da transição (isto é, a interseção obtida no agrupamento da sequência de dois intervalos consecutivos). O valor do peso utilizado para considerar duas áreas relacionadas é um parâmetro da técnica e deve ser definido de acordo com o desejo do usuário. Se é necessário conhecer todas as relações entre os tópicos, por mais fracas que sejam, uma aresta de peso baixo é considerada uma relação entre as áreas. Caso contrário, pode-se definir que somente áreas com relação forte são consideradas transições. Por exemplo, se uma área A1 transaciona para uma área A2 com uma proporção de 95% (95% dos documentos de A1 estão presentes no mesmo grupo que A2 quando ambos são agrupados juntos) e para uma área A3 com 5%, pode-se querer considerar apenas A2 como evolução de A1. Para isso, basta considerar transições evolucionárias apenas arestas com peso mínimo de 90%, o que descartaria a relação fraca entre A1 e A3.

A Figura 22 mostra o exemplo do resultado final obtido considerando-se todas as arestas transições evolucionárias (pesos maiores que 0 já são considerados transições).

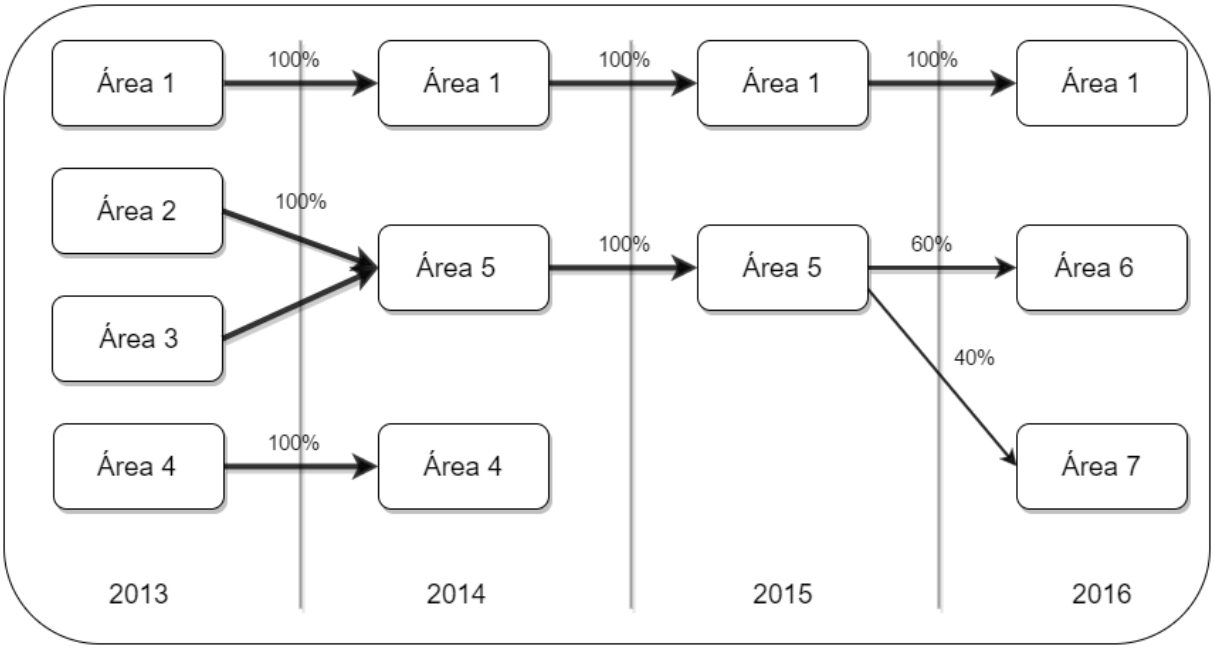



Figura 21: Exemplo de Grafo de evolução dos tópicos completo

Quadro 7: Comparativo das técnicas de análise temporal

| **Técnica** | **Características Principais** | **Vantagens** | **Desvantagens** |
|---|---|---|---|
| Algoritmos de modelagem de tópicos dinâmicos | Criam tópicos divididos no tempo ao longo da execução do algoritmo | Não é necessário nenhum processamento posterior à execução do algoritmo | Encontra somente tópicos presentes em toda a duração da coleção; tópicos do passado influenciam os tópicos do futuro; escolhas de parâmetros para cada fatia de tempo |
| Visualizações | Cria representações dos elementos da coleção e os divide temporalmente | Não necessita associar diretamente tópicos ao longo do tempo, deixa a tarefa a cargo do usuário | Difícil interpretação principalmente para grandes coleções; Exige conhecimento especializado para definir os tópicos |
| Citações entre documentos | Utiliza o próprio tempo contido nas publicações como fator de associação temporal | Mais preciso pois as citações são informações presentes nos próprios documentos | Necessita haver citações em toda a coleção; coleção deve ser “fechada”, ou seja, citantes e citados devem estar contidos nela |
| Técnica Escolhida | Divide uma coleção temporalmente e executa os algoritmos, associando os | Tópicos de cada intervalo independentes temporalmente; não é necessário escolher parâmetros do | Necessária execução em cada intervalo de tempo; pesos diferentes devem ser usados para visualizar diferentes |

| | | | |
|---|---|---|---|
| | tópicos para cada período de tempo | agrupamento; associações com pesos | características (como remificação, migrações) |

# 6 Avaliação

A proposta foi executada em quatro cenários distintos para avaliar e comparar os resultados obtidos em relação às pesquisas existentes que realizam as mesmas tarefas manualmente e à execução manual do processo. As avaliações são qualitativas devido à complexidade e natureza das tarefas envolvidas e à falta de conjuntos de teste e métodos de avaliação quantitativa na literatura presente.

Nas próximas seções são detalhadas as definições da avaliação, execução e resultados respectivamente.

## 6.1 Definições

A princípio, a proposta foi executada em cada um dos cenários, obtendo assim um conjunto de áreas de pesquisa expressadas através dos seus respectivos rótulos. Em seguida, foi elaborado um questionário que foi aplicado aos participantes com conhecimento dos cenários utilizados (no caso, Computação), visando de forma geral:

a) Comparar semânticamente os rótulos criados manualmente e os rótulos criados automaticamente pela proposta como forma de comparação das áreas.
b) Avaliar o quão bem as áreas são identificáveis e definidas através dos rótulos.

Nos dois últimos cenários, outras comparações são realizadas ainda entre os resultados da proposta e os resultados existentes em trabalhos da literatura utilizando-se os mesmos dados para avaliar:

a) Se o número de áreas encontrado automaticamente reflete o número real existente.
b) Se a evolução temporal das áreas de pesquisa encontrada automaticamente reflete a evolução encontrada na pesquisa de especialistas.

A seguir, seguem-se outras definições importantes para a avaliação.

## 6.2 Definição do Estudo Experimental

### 6.2.1 Objeto de Estudo

Rótulos gerados automaticamente para áreas e comparação entre rótulos automáticos e manuais. Áreas de pesquisa expressadas através dos rótulos.

### 6.2.2 Foco de qualidade

Eficiência do sistema mediante a opinião dos usuários quanto aos rótulos gerados para representar áreas de pesquisa denotadas por um conjunto de documentos.

### 6.2.3 Perspectiva

O estudo foi desenvolvido sob o ponto de vista e ambiente de um indivíduo que busca extrair informações de uma coleção de documentos, especificamente os temas presentes e uma forma intuitiva de reconhece-los ou se familiarizar.

### 6.2.4 Contexto

Uma coleção de documentos na área de Ciência da Computação e a técnica desenvolvida de rotulagem automática, ambas avaliadas por profissionais de Tecnologia da Informação.

Desta forma, utilizando uma notação baseada em “Objetivo, Questão e Métrica” (GQM) (SOLINGEN; BERGHOUT, 1999), temos:

**Analisar** a utilização das técnicas propostas para rotulagem automática

**Com o propósito de** avaliar a viabilidade de sua utilização a grandes coleções e o seu grau de eficiência

**Referente** ao poder de representar e expressar tópicos, comparado aos tópicos criados manualmente

**Do ponto de vista do** profissional e especialista do domínio

**No contexto de** Ciência da Computação, mas especificamente das áreas de mineração de dados, recuperação da informação e banco de dados.

Onde:

O **Objetivo** é o de verificar a efetividade do uso da proposta para identificação e representação de áreas de pesquisa.

As **Questões** incluem:

- Os rótulos identificaram e representaram bem as áreas de pesquisa?
- Rótulos automáticos são equivalentes ou cambiáveis aos manuais?

Com as **Métricas** de:

- Representatividade: o quão bem um rótulo representa um determinado conjunto de documentos identificado como pertencendo a uma área de pesquisa
- Similaridade: Uma medida de semelhança entre rótulos gerados automaticamente e rótulos manuais

## 6.3 Planejamento do Estudo Experimental

### 6.3.1 Contexto Global

Ausência de informações automáticas extraídas de quaisquer coleções de documentos científicos (como publicações, patentes, etc) de qualquer tamanho que permitam uma análise dos temas contidos, sua evolução ao longo do tempo e uma apresentação destes através de termos que os definam.

### 6.3.2 Contexto Local

O contexto local deste estudo é focado em publicações científicas na língua inglesa, mais especificamente as publicações de Mineração de Dados, Recuperação da Informação e Banco de Dados presentes nas conferências *Knowledge Discovery and Data Mining* (KDD), Special Interest Group in Information Retrieval (SIGIR) e Simpósio Brasileiro de Banco de Dados (SBBD) respectivamente. Também dos documentos presentes na coleção *Scholar Data Challenge* (SDC), que contém áreas da Ciência da Computação em geral.

### 6.3.3 Projeto Piloto

Antes da execução do estudo, foi realizado um projeto piloto com a mesma estrutura descrita neste planejamento com o autor desta dissertação. O motivo da execução com o autor foi o de poupar recursos de tempo e pessoal, visto que a execução da avaliação depende fortemente de um trabalho manual que demanda muito dos participantes e consome muito tempo.

### 6.3.4 Participantes

O critério de seleção dos participantes foi escolher pessoas envolvidas e com conhecimento nos cenários utilizados (cenários da computação) para avaliar a relevância dos rótulos automáticos e sua relação com as áreas de pesquisa correspondentes.

### 6.3.5 Treinamento

Antes de responderem às questões formuladas, os participantes foram informados sobre a finalidade da proposta e houve uma rápida explicação sobre como preenche-las (os questionários possuem explicação detalhada como pode ser visto nos apêndices A e B). Também foi dado acesso aos principais documentos de cada área para evitar equívocos na atribuição de um rótulo e liberdade para pesquisar sobre termos mais específicos.

### 6.3.6 Instrumentação

Para a realização da avaliação foi disponibilizado aos usuários os dados de todos os documentos utilizados na identificação de áreas da coleção bem como os documentos mais relevantes de cada área de acordo com a proposta.

### 6.3.7 Critérios

Utilizaram-se neste caso, critérios qualitativos. Estes foram extraídos de notas atribuídas aos rótulos de acordo com a representatividade entre o rótulo e a área e a similaridade entre o rótulo manual e o automático, ambos respectivos à mesma área.

### 6.3.8 Hipótese nula

A hipótese nula é uma afirmativa que o estudo tem como objetivo negar. A hipótese nula deste trabalho é a de que a representação de áreas através de rótulos automáticos não é possível devido a estes não serem suficientemente relacionados à área, ou seja, não é possível utilizá-los para definição da área.

Sendo:

- $\mu_h$ = qualidade da representação de áreas através de rotulagem humana
- $\mu_a$ = qualidade da representação de áreas através de rotulagem automática

Então: $H_0 = \mu_a << \mu_h$ , ou seja, a qualidade da rotulagem automática é extremamente inferior à rotulagem humana

### 6.3.9 Hipótese alternativa

A hipótese alternativa é uma afirmativa que nega a hipótese nula. Na atual pesquisa, a hipótese alternativa determina que os rótulos gerados para as áreas através do presente trabalho representam satisfatoriamente as áreas de pesquisa subjacentes.

Sendo:

- $\mu_h$ = qualidade da representação de áreas através de rotulagem humana
- $\mu_a$ = qualidade da representação de áreas através de rotulagem automática

Então: $H_1 = (\mu_a < \mu_h)$ ou $(\mu_a = \mu_h)$ ou $(\mu_a > \mu_h)$ ou $(\mu_a >> \mu_h)$

O que significa que a rotulagem automática pode ser um pouco pior, igual, melhor ou extremamente melhor que a rotulagem humana. Neste caso, para confirmar a H1, é necessário que apenas uma das opções seja verdadeira.

### 6.3.10 Variáveis independentes

Não foram colhidas informações independentes para a análise em questão em virtude da tarefa.

### 6.3.11 Variáveis dependentes

As informações dependentes fornecidas pelos participantes foram o grau de representação de um rótulo em relação à respectiva área; a atribuição de um rótulo para uma área baseado em seus termos e documentos e o grau de similaridade entre os rótulos manual e automático.

### 6.3.12 Mecanismo de análise

O meio utilizado para a avaliação foi um questionário aplicado aos participantes e que se encontra nos apêndices A e B (O primeiro é um modelo e o segundo um exemplo real).

Ele consiste basicamente nos campos de:

- Nome do participante
- Tópico: Onde há uma lista dos 10 primeiros termos mais relevantes do respectivo tópico.
- *Label* Manual: Onde é inserido pelos participantes o rótulo manual definido por cada um.
- Uma lista de rótulos para o determinado tópico

- Ao lado de cada item da lista, escalas de pontuação para as métricas de representatividade e similaridade, respectivamente.

A avaliação das variáveis dependentes da aplicação foi efetuada por meio de uma escala de 0 a 5 (exceto a atribuição de rótulo, obviamente), indicando menor e maior grau respectivamente de representação relativa a área e similaridade entre rótulos manuais e automáticos.

Então as escalas utilizadas são chamadas no questionário de "**Representação**" e "**Semelhança**" e visam capturar respectivamente a **representatividade** e **similaridade** dos rótulos

Os valores de escala significam para cada métrica:

- Representação:
  - 5 – Representa exatamente a área de pesquisa
  - 4 – Representa bem o conteúdo
  - 3 – É possível identificar a área através do rótulo
  - 2 – Relacionado, mas talvez não seja possível a identificação
  - 1 – Totalmente Inapropriado
- Semelhança:
  - 5 – Os rótulos são idênticos ou quase iguais
  - 4 – São muito parecidos em sentido
  - 3 – Relacionados
  - 2 – Possuem alguma relação, mas sentidos diferentes
  - 1 – Totalmente diferentes

Para efeitos de avaliação, os valores das métricas entre 3 e 5 são considerados satisfatórios para o que cada métrica representa.

### 6.3.13 Nomenclaturas

Conforme consta no questionário presente nos apêndices A e B, onde quer que sejam exibidos ou discutidos os resultados será utilizada a seguinte nomenclatura com respeito aos métodos de rotulagem:

a) Para diferenciar as extrações textuais e por palavras-chave, serão utilizadas as abreviações **ET1, ET2, ET3, EP1, EP2, EP3** para cada tipo de extração.
b) Para as técnicas de ranqueamento são utilizadas as siglas **tf, degtf, mdeg** e **medeg** para denotar a frequência de termos, relação grau/frequência, grau modificado do rótulo e grau modificado estendido respectivamente.
c) As técnicas **mdeg** e **medeg** serão somente utilizadas em conjunto com **ET1, EP1** e **EP2.** Devido ao fato de estas funções já contabilizarem termos contidos em outros não houve ganho e nem sentido em utilizá-las em conjunto com n-gramas.
d) Para comparação entre o uso de um único rótulo e o uso de vários será usada a notação **top-n** para designar o número de rótulos utilizados da lista final de rótulos após ranqueamento. Assim, por exemplo, top-1 designa o uso de um único rótulo enquanto top-5 os cinco primeiros rótulos respeitando a ordem. Nos presentes cenários são utilizados **top-1** e **top-3** para comparação entre rótulos únicos e múltiplos rótulos.

Combinações das técnicas são denotadas pelo uso conjunto das nomenclaturas como pode ser visto no Quadro 8.

Quadro 8: Nomenclatura das combinações de técnicas

| **Combinação** | **Seleção de Candidato** | **Ranqueamento** |
|---|---|---|
| ET1-tf, ET2-tf, ET3-tf | ET1, ET2, ET3 | tf |
| ET1-Mdeg | ET1 | mdeg |
| ET1-degtf, ET2-degtf, ET3-degtf | ET1, ET2, ET3 | degtf |
| EP1-tf, EP2-tf, EP3-tf | EP1, EP2, EP3 | tf |
| EP1-Mdeg, EP2-Mdeg | EP1, EP2 | mdeg |
| EP1-degtf, EP2-degtf, EP3-degtf | EP1, EP2, EP3 | degtf |

## 6.4 Execução da avaliação

### 6.4.1 Seleção dos participantes

Ao total, sete pessoas participaram da avaliação. Os participantes possuíam conhecimentos técnicos dos cenários utilizados porém também foi possível consultar os documentos originais e realizar pesquisas para eliminar quaisquer dúvidas em relação aos termos. O número de participantes tem por finalidade diminuir o viés individual na tarefa de rotulagem, pois a interpretação para geração dos rótulos pode ser subjetiva, assim como sua avaliação por parte de um participante.

### 6.4.2 Instrumentação

O questionário foi enviado em meio físico ou digital aos participantes para a avaliação, dependendo da preferência, para maior flexibilidade no preenchimento e retorno.

### 6.4.3 Execução da Proposta

Avaliações ocorreram em momentos distintos e em quatro cenários diferentes, utilizando-se dados diversos tanto quanto à fonte como também ao tamanho, tipo e área do conhecimento.

Basicamente a execução de todas as avaliações seguiram um mesmo roteiro. Primeiramente realizou-se uma coleta de dados, de onde se obteve a coleção a ser usada no processo de identificação de áreas. Dada a coleção de documentos, tem-se a matéria prima que será agrupada segundo a modelagem de tópicos pelo algoritmo LDA e será também definido o número de áreas ideal pela análise de estabilidade. Após a detecção das áreas presentes na coleção inicia-se o processo de rotulagem para designar termos representativos para cada uma. Ao fim, os resultados das áreas encontradas e de seus respectivos rótulos são inseridos no questionário (conforme modelo presente nos apêndices A e B) para a avaliação.

Como os detalhes de dados, execução e avaliação foram diferentes em cada cenário, são exibidos nas próximas seções os detalhes de cada um, assim como separadas as análises dos resultados por cenário.

#### *6.4.3.1 Cenário 1 – KDD*

Nesta avaliação utilizou-se como fonte de dados artigos da conferência *Knowledge Discovery and Data Mining* (KDD). Esta coleção também foi utilizada em testes para a escolha

da técnica de agrupamento devido ao trabalho de (MEI; ZHAI, 2005) mostrar a quantidade e quais áreas de pesquisa estavam salientes nos dados, facilitando uma comparação empírica entre as técnicas.

Para construir a base de dados foram extraídos os artigos cobrindo dez anos de conferência, entre os anos 2004 e 2014, utilizando-se a biblioteca digital da *Association for Computing Machinery* (ACM).

Uma etapa de pré-processamento é realizada na coleção para o uso na modelagem de tópicos. Nesta etapa, é realizada uma tokenização para decompor o texto em seus termos. Além disso, são extraídos os títulos e resumos de cada artigo e suas palavras-chave e descritores quando disponível, cada um representando um documento. Nenhum outro processamento foi utilizado nos dados para testar e mostrar a robustez dos algoritmos utilizados.

Ao total, a coleção consiste de 1483 documentos possuindo 10506 termos únicos. Dois documentos de 2014 não foram coletados devido a estarem com publicação pendente e portanto sem as informações necessárias.

Os documentos são utilizados como entrada e na fase de seleção do número de áreas, utilizou-se como mínimo e máximo de áreas 10 e 50 respectivamente. De acordo com esses parâmetros, o número ideal de áreas encontrado foi de 37, as quais foram utilizadas na geração de rótulos.

Os parâmetros utilizados no algoritmo de geração de rótulos foram D = W = 10. Assim, uma amostra de dez documentos de cada área foi usada no processo, assim como as dez palavras mais relevantes da área.

#### *6.4.3.2 Cenário 2 – SDC*

Nesta avaliação foram utilizados dados provenientes do *Scholar Data Challenge* (SDC) (TANG et al., 2007). Essa base de dados contém documentos provenientes de diversas fontes e de diversos tipos, como artigos, livros e patentes.

A coleção contém 2.092.256 documentos e foi utilizada para avaliar os resultados com grandes volumes de dados em um cenário de *big data*. A robustez dos algoritmos e do processo como um todo também é avaliada nesse caso, visto que muitas vezes processamentos de linguagem mais complexos são muito custosos e impossíveis de executar com esta quantidade massiva de dados.

Uma etapa de pré-processamento também foi realizada neste caso. Os documentos, apesar da base ser bem heterogênea, podiam possuir uma série de campos, como título, autor, ano, editora entre outros. Primeiramente foram selecionados todos os documentos que possuíam um título e um resumo. Observando os dados presentes após essa primeira etapa, se constatou que existiam muitos documentos com resumos contendo pouca informação ou duplicatas dos respectivos títulos. Assim, para evitar que a modelagem de tópicos fosse influenciada pela presença de documentos com poucos termos (lembrando que a modelagem funciona criando distribuição de termos pelas áreas), estes foram filtrados da coleção. Após o processo de tokenização, todos os documentos que contivessem menos que 50 termos foram removidos do conjunto final.

A coleção final utilizada, após o pré-processamento, ficou com 1.057.791 documentos, cada documento consistindo de um título e resumo. A base de dados não continha palavras-chave e descritores dos dados, então, as técnicas que utilizam seleção de candidatos por palavra-chave não foram avaliadas contra esta base de dados.

Devido ao tamanho da coleção, foram extraídas 100 áreas da coleção como forma de possuir temas mais gerais e possíveis de serem avaliados por pessoas.

Os parâmetros D e W utilizados aqui foram os mesmos usados no primeiro cenário e ambos iguais a dez.

#### *6.4.3.3 Cenário 3 – SIGIR*

Este cenário consistiu em dados sobre artigos oriundos da conferência SIGIR (conferência em armazenamento e recuperação da informação). A escolha desse cenário e base para a avaliação é devida aos trabalhos de (SMEATON et al., 2002), que realiza uma análise de 25 anos da conferência investigando os principais tópicos e temas de pesquisa da área e de (TSENG et al., 2009), que aborda a evolução dos temas ao longo do tempo para identificar áreas “quentes” (com interesse crescente). Ambos os trabalhos fazem análises totalmente manuais para as respectivas tarefas em cima da mesma base deste cenário.

O objetivo deste cenário é, além de avaliar a atribuição de rótulos e a consistência das áreas, comparar os resultados com os trabalhos manuais existentes. Adicionalmente será avaliado se a análise temporal proposta é correspondente aos dados existentes nas pesquisas (ambos os trabalhos realizam uma análise temporal da conferência).

A coleção original consiste nos artigos aceitos ao longo de 25 anos da conferência SIGIR, contando a partir da primeira até a vigésima quinta em ordem de tempo. Assim, os dados abrangem desde a primeira edição em 1971 até a de 2002. Uma observação é que depois da edição de 1971 a próxima edição só ocorreu em 1978, por isso são 25 edições e não 31. Como entre a primeira edição e a segunda constante na base há uma diferença de 6 anos (1971-1978) não será considerada a primeira edição para a análise temporal, neste caso começando de 1978 até 2002.

A base completa possui 853 documentos, os quais foram extraídos a partir da biblioteca digital da ACM (como no primeiro cenário de avaliação). Título e resumo foram coletados e transformados na coleção de entrada. Os documentos também foram divididos por edição para a análise temporal (visto que cada edição corresponde a um ano). Somente tokenização foi realizada no texto.

Não estão presentes palavras-chave e classificadores na coleção, deste modo as técnicas de rotulagem que fazem uso destas informações não serão avaliadas aqui.

O trabalho original (SMEATON et al., 2002) identificou 29 áreas no total (nos 25 anos) e atribuiu um rótulo para cada uma delas. Depois, os documentos de cada área foram ligados aos respectivos anos gerando uma distribuição dos temas ao longo do tempo. De acordo com os autores, o número mínimo de áreas ao longo de todas as edições foi 5 e o máximo 20. Ao passar pela seleção de áreas desta proposta se utilizou então os parâmetros 4 e 25 como números de áreas mínimo e máximo (parâmetros para seleção de áreas) a cada ano. Quando houveram mais de um número ótimo de áreas em diferentes granularidades optou-se por escolher o valor mais próximo ao utilizado pelos autores para efeito de comparação.

Incluindo áreas que se repetem ao longo dos anos, o trabalho fonte possui ao longo das edições 336 áreas (29 áreas distintas com o resto sendo duplicatas de áreas existentes em anos anteriores). Devido ao grande número de áreas idênticas em diferentes anos, a avaliação apresentou aos participantes somente áreas novas no questionário. Áreas que possuíram uma transição evolucionária com um peso mínimo de 0,6 foram consideradas a mesma área e por isso foram rotuladas pelos participantes apenas uma vez.

Os parâmetros para a rotulagem D e W foram iguais a dez, porém algumas áreas em determinados anos possuíam menos de dez documentos o que poderia de alguma forma influenciar os resultados nestes casos.

#### *6.4.3.4 Cenário 4 – SBBD*

A coleção deste último cenário foi baseada no trabalho de (KAUER, 2013), que analisa a evolução dos temas mais relevantes ao longo dos anos na conferência SBBD (Simpósio Brasileiro de Banco de Dados). Os autores realizam um agrupamento nos artigos presentes nas edições entre 1986 e 2012 e definem empiricamente as áreas mais relevantes ao longo das edições. Também são atribuídos rótulos manuais para os temas de pesquisa conforme os autores.

O objetivo aqui também é comparar os resultados automáticos obtidos aos manuais dos autores em questão. Como também é realizada uma análise temporal no artigo original, é possível uma comparação dos resultados temporais.

A coleção original possui 475 documentos contendo o resumo dos artigos. As edições de 1986 a 1988 foram descartadas pois os autores utilizaram apenas artigos com resumo em inglês.

Para essa avaliação foram utilizados apenas os artigos presentes nas edições de 1999 a 2012. O motivo é devido ao fato de as edições anteriores não estarem em formato digital e portanto não disponíveis na internet, exigindo uma obtenção manual dos dados. O fato de utilizar esse subconjunto não fere a comparação entre os resultados (é possível comparar os dados apenas de 1999 a 2012), então optou-se por utilizar apenas os dados disponíveis. Adicionalmente, pode-se avaliar também o desempenha da proposta em uma coleção pequena de dados, visto que é a menor coleção entre os cenários. Todos os dados foram extraídos da base dblp[1] e lbd-ufmg[2].

Nem todos os documentos possuíam descritores e palavras-chave, por isso não foram utilizados para a rotulagem já que poderiam faltar muitos dados. Os parâmetros D e W da rotulagem foram iguais a dez, porém também vale notar que alguns grupos possuíam menos que dez documentos.

A coleção foi dividida por ano (edição) e novamente o único pré-processamento envolvido foi a tokenização. No artigo original o número de áreas varia ano a ano e fica entre três e seis dependendo da edição. Baseando-se nisso foi utilizado como parâmetro para a

[1] Disponível em: http://dblp.uni-trier.de/

[2] Disponível em: http://www.lbd.dcc.ufmg.br/

seleção de áreas os valores mínimo e máximo de 3 e 8 respectivamente. O número de áreas varia e caso haja duas possibilidades para o número ótimo de áreas escolhe-se a que estiver mais próxima a granularidade usada pelos autores.

No trabalho fonte, são analisadas as ramificações e fusões de diferentes temas de pesquisa ao longo do tempo. Por esse motivo, aqui serão consideradas as transições evolucionárias com peso maior ou igual a 0,4 (serão exibidas no grafo as arestas que respeitam essa restrição). Com restrições de peso maiores, não seria possível visualizar as ramificações por exemplo, porque o somatório dos pesos é sempre igual a um. Isso também pode acarretar o aparecimento de arestas que não constam na fonte como efeito colateral.

### 6.4.4 Execução do Questionário

A todos os participantes foi dado um questionário (presente nos apêndices A e B) para avaliar qualitativamente os rótulos. As áreas foram representadas no questionário pelos seus dez termos mais relevantes segundo a modelagem de tópicos. Em seguida, são apresentados os rótulos automáticos criados utilizando várias combinações de técnicas conforme definido nas nomenclaturas. Juntamente com os rótulos automáticos foi inserido um rótulo consistindo dos cinco termos mais relevantes de acordo com o LDA para efeito de comparação com os métodos tradicionais de interpretação de tópicos que utilizam massivamente essa lista de termos.

As seguintes tarefas foram realizadas pelos participantes:

1. Atribuir um rótulo manual para a respectiva área. Lembrando que foi possível pesquisar e consultar os documentos em caso de dúvida. Para os cenários 3 e 4 (SIGIR e SBBD), ao invés de atribuir um rótulo foi pedido para selecionar um rótulo de uma lista. A lista consiste nos rótulos já presentes nos trabalhos nos quais os respectivos cenários são baseados com a adição de uma opção “outros” para documentos que não se encaixem em nenhum rótulo.
2. Dar uma nota de um a cinco quanto a representação. Essa pontuação define o quão bem um rótulo está representando a área. Uma nota um, por exemplo, significa que o rótulo é totalmente inapropriado enquanto uma nota cinco significa que o rótulo expressa exatamente a temática da área. Pontuações de três a cinco são consideradas satisfatórias nesse caso, já que é possível descobrir o tema com rótulos que tenham pontuação nesse intervalo.

3. Dar uma nota de um a cinco de acordo com o quão semanticamente similar o rótulo avaliado é ao rótulo atribuído pelo participante (Medida de similaridade). Da mesma forma que a nota de representação, uma pontuação de três a cinco é considerada satisfatória, ou seja, existe uma relação próxima entre os rótulos. Uma pontuação de cinco significa que os rótulos são idênticos enquanto a pontuação mínima indica que não nenhuma relação entre eles.

A medida de representação reflete então a qualidade do rótulo em representar o conteúdo da área. A medida de similaridade ajuda a avaliar algumas limitações que são usualmente encontradas quando se realiza uma avaliação quantitativa de rotulagem, como por exemplo tentar comparar se rótulos são lexicamente idênticos. Nestes casos, qualquer rótulo que tenha um sinônimo ou que seja semanticamente idêntico terá sempre uma pontuação menor. Essa medida pode ajudar a avaliar se rótulos distintos podem ambos ser usados satisfatoriamente.

## 6.5 Resultados

Nesta seção são apresentados os resultados da execução e avaliação também divididos por cenário (nas próximas subseções) para facilitar a compreensão dos dados. A nomenclatura utilizada para visualizar o resultado dos métodos é a mesma apresentada na definição da avaliação. Resultados completos da execução da proposta podem ser visualizados nos Apêndices C, D, E e F (respectivos aos cenários 1, 2, 3 e 4).

### 6.5.1 Cenário 1 – KDD

O Quadro 9 mostra um exemplo das áreas agrupadas pela modelagem de tópicos juntamente com as 10 palavras mais relevantes associadas com cada uma. É importante notar aqui que embora algumas sejam de fácil compreensão (por exemplo, a área 2 é sobre métodos de otimização), nem sempre é fácil deduzir o assunto ou defini-las pelos conjuntos de termos, especialmente para alguém que não esteja no contexto de cada uma ou que não tenha maiores informações sobre o domínio.

Quadro 9: Exemplos de Áreas detectadas

| Área 1 | Área 2 | Área 3 | Área 4 |
|---|---|---|---|

| | | | |
|---|---|---|---|
| identify | optimization | knowledge | system |
| disease | methods | accuracy | mining |
| medical | proposed | detection | management |
| identifying | formulation | sample | techniques |
| health | show | available | analysis |
| study | functions | standard | systems |
| records | solve | given | application |
| clinical | linear | requires | large |
| features | regression | work | designed |
| patients | propose | performance | high |

A seguir, no Quadro 10, são exibidos os resultados dos variados métodos (incluindo as 5 primeiras palavras do tópico) e seus respectivos rótulos para algumas das áreas. A primeira linha contém os rótulos manuais atribuídos pelos participantes (onde a maioria concordou com o rótulo).

Quadro 10: Áreas com seus respectivos rótulos gerados

| **Técnica** | **Rótulos** | | | |
|---|---|---|---|---|
| **Manual** | **social networks** | **clustering** | **active learning** | **recommender systems** |
| ET1-TF | social networks, nodes, connection subgraphs | clusters, clustering, algorithms | active learning, labeled data, labels | collaborative filtering, recommender systems, users |
| ET2-TF | social networks, large social network graphs, nodes | clusters, clustering, algorithms | labeled data, labels, unlabeled data | collaborative filtering, recommender systems, users |
| ET3-TF | social networks, large social network graphs, networks graphs | clusters, clustering, algorithms | labeled data, active learning, labels | collaborative filtering, recommender systems, users |
| ET1-MDeg | social networks, social network, graph | clustering, clusters, subspace cluster | active learning, label, labeled data | collaborative filtering, recommender systems, users |
| ET1-DTF | social network, large network, data structures | real data, categorical objects, | binary classification, active | tag recommender, recommendation based, |

| | | | | |
|---|---|---|---|---|
| | | subspace clustering | learning, active labeling | recommender systems |
| ET2-DTF | large network, social network, large networks | real data, categorical objects, real world | classification algorithm, binary classification, learning algorithm | tag recommender, recommendation based, filtering methods |
| ET3-DTF | social communication, communication network, compressing social | real data, quality hierarchical, approach seamlessly | fully supervised, classification algorithm, label efficient | user preferences, user posts, reputable user |
| EP1-TF | social networks | clustering | active learning | recommender systems |
| EP2-TF | social networks | clustering | learning | recommendation |
| EP3-TF | networks | data mining | active learning | data mining |
| EP1-MDeg | quality | clustering | data mining | subspace clustering |
| EP2-MDeg | experimentation | episode mining | information search and retrieval | learning |
| EP1-DTF | user generated content | minimum description length principle | interactive and online data mining | hybrid content and collaborative filtering |
| EP2-DTF | user generated content | minimum description length principle | interactive and online data mining | user profiles and alert services |
| EP3-DTF | learning | kernel | Misclassification | filtering |
| Top-5 palavras | graph, graphs, network, nodes, networks | clustering, cluster, clusters, objects, experiments | training, labeled, classification, classifier, supervised | users, recommendation, user, system, collaborative |

Aqui são apresentados os três primeiros rótulos para uma melhor comparação entre a abordagem com um único rótulo (top-1) e de múltiplos rótulos (top-3). Claramente, na abordagem de um único rótulo exibe-se somente o primeiro dos três pois estão ordenados da mesma forma como saem do ranqueamento. As abordagens que utilizam palavras-chave e descritores (EP1, EP2, EP3) não possuem mais de um rótulo devido à escassez de termos.

Por fim, é apresentado no Quadro 11 o resultado das avaliações usando-se a média de pontuações para cada métrica utilizada nas medidas de representatividade e similaridade.

Quadro 11: Média das pontuações para cada técnica de rotulagem utilizada (KDD)

| **Medida** | **Representação** | | **Similaridade** | |
|---|---|---|---|---|
| **Número de rótulos / Método** | **(top-1)** | **(top-3)** | **(top-1)** | **(top-3)** |
| **ET1-TF** | 3.69 | 3.83 | 3.15 | 3.53 |
| **ET2-TF** | 3.34 | 3.50 | 3.08 | 3.17 |
| **ET3-TF** | 3.22 | 3.45 | 3.01 | 3.32 |
| **ET1-MDeg** | 3.75 | 3.85 | 3.13 | 3.55 |
| **ET1-DTF** | 2.64 | 3.52 | 2.30 | 2.36 |
| **ET2-DTF** | 2.58 | 3.48 | 2.27 | 2.36 |
| **ET3-DTF** | 2.54 | 3.36 | 2.07 | 2.34 |
| **EP1-TF** | 3.45 | - | 2.87 | - |
| **EP2-TF** | 2.71 | - | 2.03 | - |
| **EP3-TF** | 2.60 | - | 1.89 | - |
| **EP1-MDeg** | 2.54 | - | 1.96 | - |
| **EP2-MDeg** | 2.12 | - | 1.25 | - |
| **EP1-DTF** | 1.72 | - | 1.37 | - |
| **EP2-DTF** | 1.35 | - | 1.30 | - |
| **EP3-DTF** | 1.22 | - | 1.19 | - |
| **Top-5** | 2.02 | - | 1.31 | - |

### 6.5.2 Cenário 2 – SDC

Para a base de dados *Scholar Data Challenge*, o Quadro 12 mostra algumas das áreas identificadas como exemplo.

Quadro 12: Exemplos de áreas detectadas (SDC)

| **Área 1** | **Área 2** | **Área 3** | **Área 4** |
|---|---|---|---|

| | | | |
|---|---|---|---|
| query<br>queries<br>database databases<br>relational<br>processing<br>xml<br>efficient<br>schema<br>querying | network<br>networks<br>services<br>mobile<br>service<br>internet<br>wireless<br>access<br>multimedia<br>ip | students<br>university<br>course<br>education<br>science<br>teaching<br>computer<br>learning<br>courses<br>engineering | signal<br>noise<br>filter<br>signals<br>estimation<br>frequency<br>filters<br>transform<br>filtering<br>linear |

Já no Quadro 13 são apresentadas as médias das pontuações da avaliação. Como a base de dados deste cenário não possuía palavras-chave e descritores, apenas as técnicas puramente textuais foram avaliadas, por isso a ausência de EP1, EP2 e EP3.

Quadro 13: Média das pontuações para cada técnica de rotulagem utilizada (SDC)

| **Medida** | **Representação** | | **Similaridade** | |
|---|---|---|---|---|
| **Número de rótulos / Método** | **(top-1)** | **(top-3)** | **(top-1)** | **(top-3)** |
| **ET1-tf** | 3.13 | 3.30 | 3.08 | 3.23 |
| **ET2-tf** | 3.03 | 3.28 | 2.85 | 2.99 |
| **ET3-tf** | 3.02 | 3.11 | 2.78 | 2.97 |
| **ET1-mdeg** | 3.25 | 3.31 | 3.05 | 3.34 |
| **ET1-medeg** | 3.43 | 3.64 | 3.11 | 3.40 |
| **ET1-deg/tf** | 2.49 | 2.55 | 2.09 | 2.11 |
| **ET2-deg/tf** | 2.45 | 2.51 | 2.06 | 2.07 |
| **ET3-deg/tf** | 2.07 | 2.12 | 2.01 | 2.01 |
| **Top-5** | 2.33 | | 2.07 | |

### 6.5.3 Cenário 3 – SIGIR

O Quadro 14 mostra alguns rótulos e a comparação com os rótulos presentes em (SMEATON et al., 2002) como exemplo.

Quadro 14: Exemplos de rótulos gerados para algumas áreas (SIGIR)

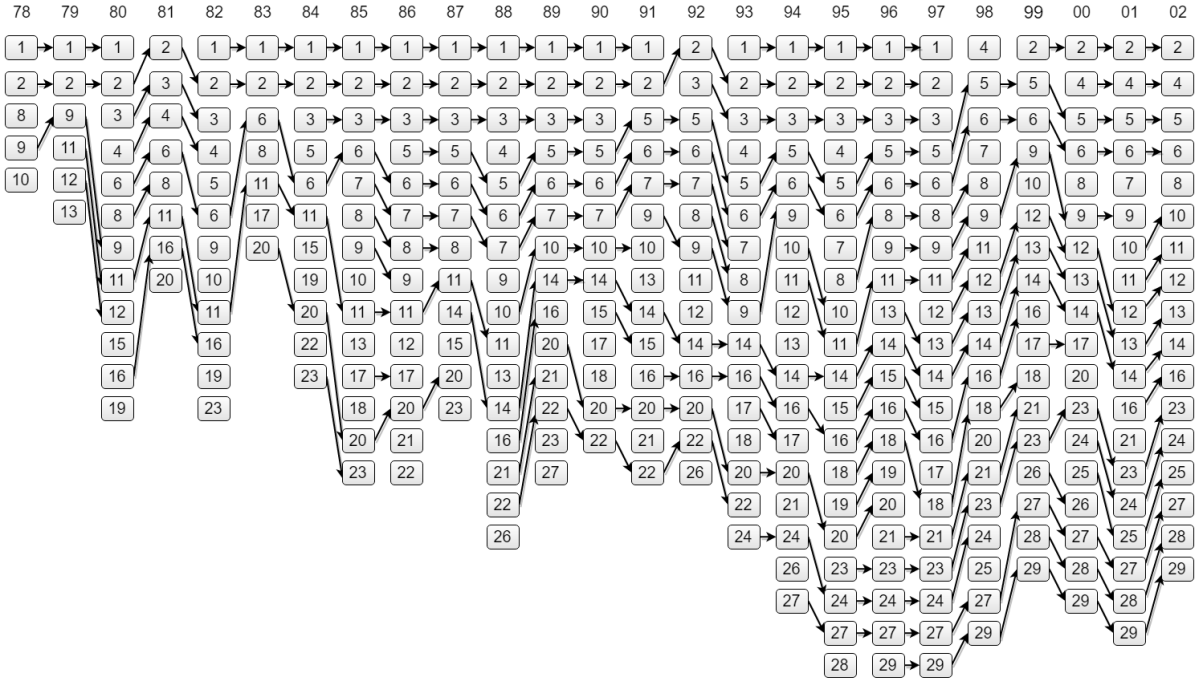
| **Técnica/Rótulo** | | |
|---|---|---|
| **Manual** | Visualization | Text Categorisation |
| **ET1-tf** | Visualizations, systems, representation | Classification, textual, linear |
| **ET2-tf** | Visualizations, systems, model | Classification, textual, algorithm |
| **ET3-tf** | Visualizations, model, information | Classification, textual, algorithm |
| **ET1-mdeg** | Visualizations, information, information representation | Text classification, textual, algorithm |
| **ET1-medeg** | Visualizations, information representation, graphic model | Classification, classification algorithm, textual data |
| **ET1-deg/tf** | Data flow, visualizations, data analysis | Textual data used, supervised approaches, algorithm |
| **ET2-deg/tf** | Data flow, analysis, data visualization | Textual data, classification algorithm, supervised approaches |
| **ET3-deg/tf** | Graphic usage, data analysis, data | Textual data used, supervised approaches include, main classification algorithm |
| **Top-5** | Data, visual, system, information, graphic, representation | Approach, classification, main, text, algorithm |

Já a Figura 22 mostra uma comparação entre as relações temporais presentes no artigo original e as encontradas pela proposta. Os rótulos exibidos para a proposta são baseados na

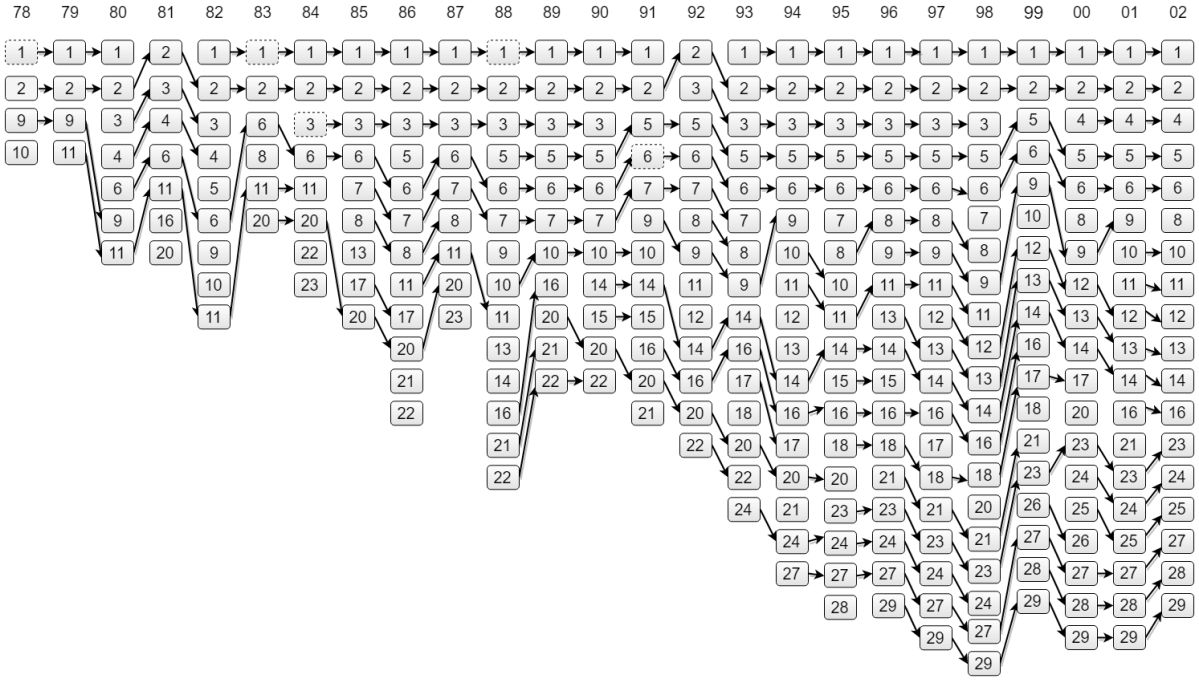
maioria das atribuições dos participantes. Os temas foram dispostos de forma a facilitar o acompanhamento das áreas ao longo do tempo e a comparação entre os dois resultados.

A área número 2 (“Geral”) corresponde aos documentos que não se encaixaram em nenhum dos rótulos, como uma miscelânea. Esta área é a única presente em todas as edições. Seguem de perto as áreas 1 (Banco de dados, Interfaces de Linguagem Natural) e 3 (Modelos) que também correspondem a tópicos muito abrangentes. A área 19 (Sistemas Gerenciadores de Bancos de dados e Recuperação da Informação) não foi encontrada ou atribuída a nenhuma das áreas extraídos pela proposta.

Nos casos de ocorrência de duas áreas no mesmo ano com o mesmo rótulo atribuído, foi realizada a integração destas e de suas arestas com os respectivos pesos calculados como uma média simples.


78 79 80 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99 00 01 02

(a)


78 79 80 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99 00 01 02

(b)

Legenda:

Ocorrência de dois grupos que receberam o mesmo rótulo

1. Banco de Dados, Interfaces de Linguagem Natural
2. Geral
3. Modelos
4. Resposta à questões
5. Frases Sintáticas & recuperação de documentos de áudio
6. Recuperação da Informação Conceitual, Bases de conhecimento em Recuperação da Informação
7. Compressão
8. Clusterização
9. Feedback de relevância
10. Arquivos Invertidos & implementações
11. Peso de termos
12. Entendimento de Mensagens & Rastreio de Eventos
13. Filtragem
14. Recuperação da Informação via Hipertexto, evidências múltiplas
15. Recuperação de imagens
16. Modelos linguísticos e probabilísticos
17. Booleanos & booleanos extendidos
18. Recuperação da informação em chinês e japonês
19. Sistemas Gerenciadores de bancos de dados e Recuperação da informação
20. Usuários & Pesquisa
21. Visualização
22. Arquivos de assinatura
23. Recuperação da informação distribuída
24. Avaliação
25. Recuperação de ligações
26. Indexação Semântica Latente
27. Classificação Textual
28. Sumarização de documentos
29. Multi-idiomas

Figura 22: Grafo de evolução dos tópicos de (a) trabalho fonte (SMEATON et al., 2002) e (b) proposta.

Por fim, o Quadro 15 mostra a média das pontuações da avaliação neste cenário.

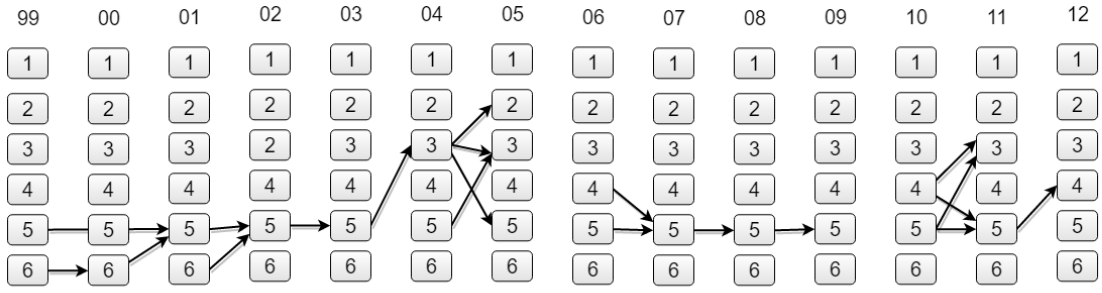
Quadro 15: Média das pontuações para cada técnica de rotulagem utilizada (SIGIR)

| **Medida** | **Representação** | | **Similaridade** | |
|---|---|---|---|---|
| **Número de rótulos / Método** | **(top-1)** | **(top-3)** | **(top-1)** | **(top-3)** |
| **ET1-tf** | 3.24 | 3.90 | 2.58 | 2.21 |
| **ET2-tf** | 3.11 | 3.61 | 2.84 | 2.83 |
| **ET3-tf** | 3.25 | 3.85 | 2.73 | 2.78 |
| **ET1-mdeg** | 3.90 | 3.96 | 2.84 | 2.80 |
| **ET1-medeg** | 3.87 | 3.98 | 2.18 | 2.51 |
| **ET1-deg/tf** | 2.08 | 2.75 | 2.06 | 2.03 |
| **ET2-deg/tf** | 2.18 | 2.56 | 2.14 | 2.89 |
| **ET3-deg/tf** | 2.38 | 2.87 | 2.57 | 2.50 |
| **Top-5** | 2.50 | | 2.55 | |

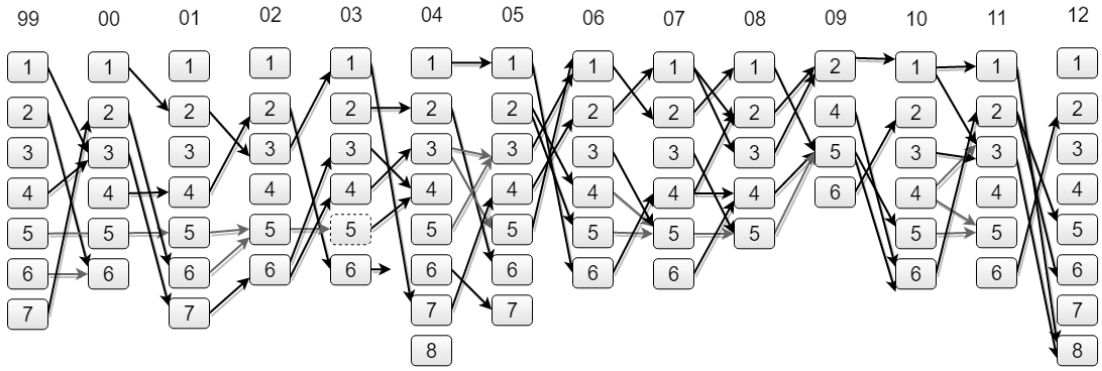
### 6.5.4 Cenário 4 – SBBD

A Figura 23 exibe uma comparação entre as evoluções temporais inferidas por (KAUER, 2013) e as encontradas pela proposta.

Uma observação acerca dos grafos é que no trabalho fonte, alguns rótulos não foram nomeados pelos autores, sendo designados por “tema 1”, “tema 2”, etc. Deste modo, esses casos não permitem uma comparação direta. Outro fato é que os autores parecem focar mais na transição entre áreas aparentemente diferentes, pois não há arestas entre áreas com rótulos idênticos ou semelhantes. Um exemplo são as áreas 1 de 2000 e 2 de 2001, ambas designadas como “dados temporais” mas sem nenhuma ligação no respectivo grafo.




(a)




(b)

Legenda:

1999:

1. Modelos multi-dimensionais
2. Banco de dados distribuídos
3. Bases de Conhecimento
4. Metadados
5. Técnicas para criação de índices
6. Persistência
7. Dados Heterogêneos

2000:

1. Dados temporais
2. Transferência de dados
3. Data warehouse
4. Processamento e recuperação de e-mails
5. Restrições de integridade
6. Processamento paralelo de consultas

2001:

1. Técnicas para criação de índices
2. Dados temporais
3. Data mining
4. Recuperação da informaçao
5. integração de dados
6. versionamento de esquemas
7. Data warehouses

2002:

1. Criptografia de dados
2. Performance de consultas
3. Dados semi-estruturados
4. Banco de dados orientado a objetos
5. Recuperação de informação multilíngue
6. Ferramentas de auxílio à tomada de decisão

2003:

1. XML
2. Sistemas distribuídos
3. Data Warehouse
4. Data Mining
5. Recuperação e integração de dados
6. Consultas

2004:

1. Dados Web
2. Sistemas distribuídos
3. Data mining
4. Data Warehouse
5. Dados temporais
6. Processamento de consultas
7. XML
8. Outros

2005:

1. Sistemas Web
2. Similaridade de objetos
3. Mineração de dados temporais
4. XML
5. Regras de associação e classificação
6. Peer-to-Peer
7. SQL

2006:

1. Data Mining
2. XML
3. Modelos multidimensionais
4. Versionamento
5. Representação de Objetos
6. Web

2007:

1. XML
2. Mineração
3. Similaridade
4. Bibliotecas digitais
5. Bancos de dados geo-temporais
6. Sistemas distribuídos

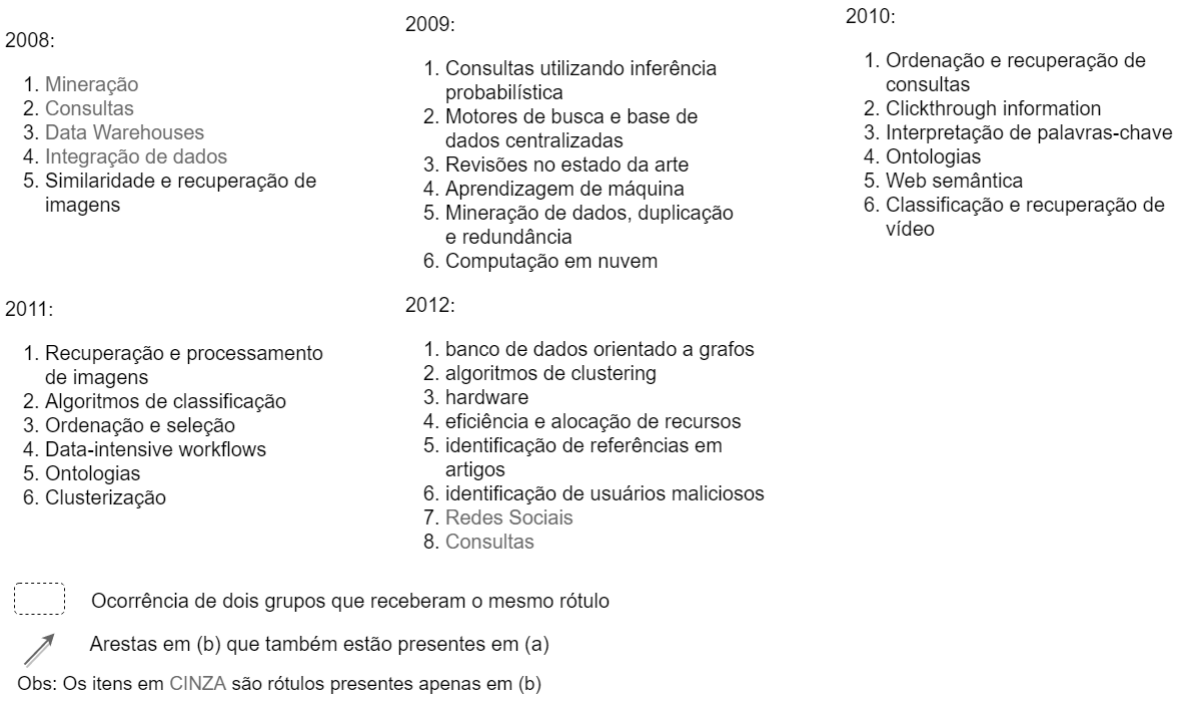



Figura 23: Grafo de evolução dos tópicos de (a) trabalho fonte (Adaptado de (KAUER, 2013)) e (b) proposta

Conforme os outros cenários, o Quadro 16 mostra a média das pontuações da presente avaliação.

Quadro 16: Média das pontuações para cada técnica de rotulagem utilizada (SBBD)

| **Medida** | **Representação** | | **Similaridade** | |
|---|---|---|---|---|
| **Número de rótulos / Método** | **(top-1)** | **(top-3)** | **(top-1)** | **(top-3)** |
| **ET1-tf** | 3.75 | 3.79 | 3.13 | 3.56 |
| **ET2-tf** | 3.22 | 3.23 | 3.10 | 3.20 |
| **ET3-tf** | 3.55 | 3.58 | 3.19 | 3.16 |
| **ET1-mdeg** | 3.64 | 3.66 | 3.08 | 3.43 |
| **ET1-medeg** | 3.88 | 4.03 | 3.89 | 3.93 |
| **ET1-deg/tf** | 3.37 | 3.68 | 2.67 | 3.00 |
| **ET2-deg/tf** | 2.78 | 2.23 | 2.33 | 2.64 |
| **ET3-deg/tf** | 2.31 | 2.87 | 2.02 | 2.64 |
| **Top-5** | 3.13 | | 2.78 | |

### 6.5.5 Análise dos Resultados

Em geral, os resultados mostram notas satisfatórias quando usadas algumas técnicas e pontuações ruins com o uso de outras combinações.

A métrica de ranqueamento deg/tf mostrou-se consistentemente pior em relação as outras e, em média, abaixo do nível considerado “bom”. As melhores técnicas de ranqueamento foram tf, mdeg e medeg com pontuações parecidas, porém mdeg e medeg aparecem acima da tf em praticamente todas os casos.

Entre as técnicas baseadas em grau do rótulo (mdeg e medeg), houve um pequeno ganho de pontuação ao usar a última. Entretanto, ambas não tiveram uma boa performance quando usadas em conjunto com extrações de candidatos por palavras-chave. Neste caso, a tf obteve um desempenho melhor, indicando que é melhor utilizada quando os candidatos já são palavras-chave enquanto as outras são mais indicadas para termos extraídos puramente do texto. Uma explicação para essa discrepância pode ser de que não há muito ganho em utilizar essas técnicas com rótulos formados por palavras-chave, pois estas costumam ser em sua maioria bigramas ou n-gramas.

A dificuldade em se interpretar os tópicos por listas de termos também ficou evidente pela pontuação obtida pelo top-5, que consistia das cinco palavras mais relevantes da área. Ele obteve as pontuações mais baixas quando comparados com a maioria das técnicas avaliadas.

Entre as variações na seleção de candidatos, não houveram diferenças significativas entre ET1, ET2 e ET3 e nem mesmo entre EP1, EP2 e EP3. A adição de mais bigramas e trigramas não causou impacto significativo nos resultados.

Já no número de rótulos, houve aumentos consistentes nas pontuações das medidas de representividade e similaridade quando usada uma abordagem de múltiplos rótulos (top-3). A diferença pode ser visualizada na Figura 24.

Todos as técnicas apresentaram também pontuações em relação à similaridade menores do que as pontuações de representatividade. Isso é uma evidência de que mesmo que um rótulo não seja similar ao rótulo manual, ele ainda sim pode ser usado para representar bem os conceitos da área.

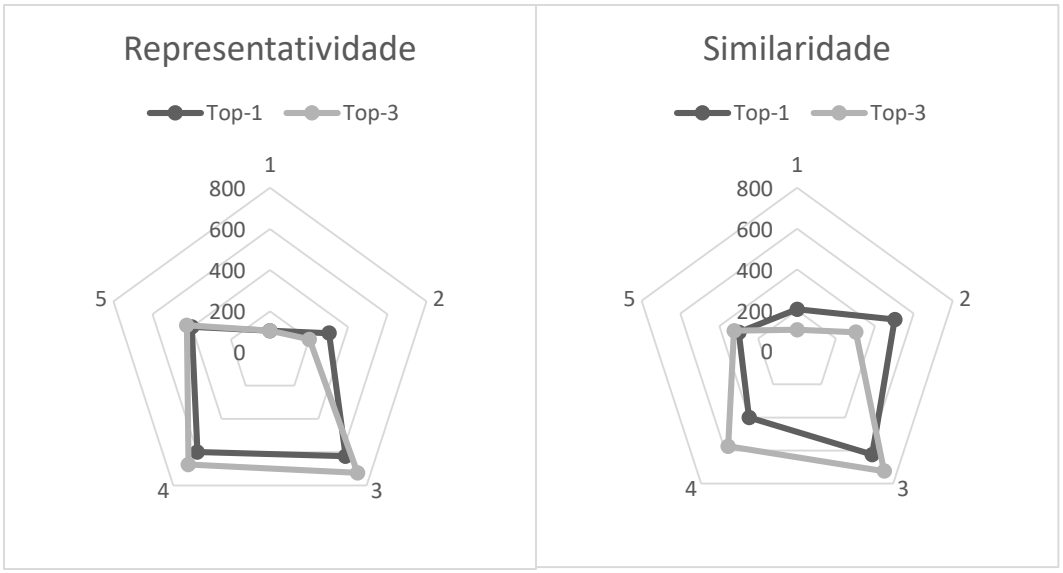



Figura 24: Diferenças nas pontuações entre o uso de rótulos únicos (top-1) e múltiplos (top-3)

Resumindo, os resultados evidenciam que ET1-tf, ET1-mdeg e ET1-medeg são as melhores combinações para uso quando utilizado somente o conteúdo textual dos documentos com uma pequena vantagem para ET1-medeg. Quando utilizados múltiplos rótulos, ET1-medeg tem uma vantagem mais robusta em relação as outras, o que pode ser justificado pela preferência da técnica por frases curtas e sucintas.

No caso de já haver descritores ou palavras-chave contidas na coleção a combinação EP1-tf foi a melhor escolha. Talvez os descritores presentes foram genéricos demais para representar a área em relação as palavras-chave definidas pelos autores.

Desta forma, pode-se dizer que pelo menos nos casos de ET1-tf, ET1-mdeg, ET1-medeg e EP1-tf os rótulos gerados são suficientes para representar as áreas. Considerando que as pontuações acima de três são equivalentes a rotulagens manuais, também pode-se dizer que estes rótulos são muito semelhantes aos rótulos manuais em termos de representatividade.

#### *6.5.5.1 Comparação entre cenários*

Outro fato interessante é que a base de dados KDD obteve pontuações melhores do que as obtidas pela base SDC para as tarefas de identificar e rotular as áreas de pesquisa. Fatores que podem explicar esse resultado podem ser a quantidade de áreas extraídas ou até mesmo a natureza da coleção. Como a base SDC teve menos áreas extraídas em relação ao total de documentos da coleção isso pode ter tornado as áreas detectadas mais gerais, ou

seja, com um grão maior. Outro fator é que a base SDC envolve um número variado de fontes, tendo uma natureza mais ampla e abrindo mais interpretações acerca de uma área específica.

Já entre as coleções SIGIR e SBBD, há claramente pontuações bem melhores para a última. Um motivo para essa diferença pode ser a forma como as pesquisas originais dos respectivos autores foram conduzidas (Os rótulos manuais nesses dois casos consistiam nos rótulos atribuídos por (KAUER, 2013; SMEATON et al., 2002) respectivamente). Um comentário recorrente dos participantes foi o de que apesar dos rótulos gerados serem muitas vezes bons, os rótulos definidos no trabalho original de (SMEATON et al., 2002) eram muitas vezes genéricos demais, apesar de servirem ao propósito. Alguns dos rótulos dados pelos autores incluíam termos genéricos para a área de recuperação da informação como “Banco de Dados” e “Modelos” ou mesmo agrupavam duas áreas em um único rótulo, como “Frases Sintáticas & recuperação de documentos de áudio”. Esses fatores podem ter tido influência principalmente nas notas de similaridade que buscam saber se um rótulo gerado é semanticamente semelhante ao manual ou existente. Outro fator que pode ter contribuído para o resultado é a esparsidade da coleção em relação aos grupos. Muitas áreas das 29 originais possuíam apenas um documento em determinados anos na pesquisa original sobre a SIGIR. Muitas delas em determinados anos podem não ter aparecido na execução da proposta devido ao fato de não estarem salientes ou relevantes dentre os dados da coleção.

Em (KAUER, 2013), os rótulos atribuídos às áreas já são mais específicos e os próprios temas de pesquisa mais densos, o que pode ter facilitado a similaridade entre os rótulos gerados e os manuais.

Uma tendência pode ser evidenciada por estes resultados combinados aos resultados da rotulagem com o uso de descritores. Quanto maior o nível de abstração dos rótulos, sejam manuais ou automáticos, as notas de representação tendem a cair. Essa evidência sugere que apesar de termos genéricos poderem ser usadas na rotulagem os usuários preferem termos mais “precisos”. Termos mais específicos poderiam ajudar a entender melhor os grupos encontrados.

#### *6.5.5.2 Comparação de áreas*

O agrupamento é uma tarefa não-supervisionada e portanto na grande maioria dos casos não há um padrão ou conjunto verdade (*gold standard*) para efeitos de comparação (CHANG et al., 2009).

Para relacionar as áreas encontradas foram utilizadas as bases SIGIR e SBBD que possuíam um número específico de áreas previamente rotuladas pelos autores (SMEATON et al., 2002) e (KAUER, 2013) respectivamente.

Utilizou-se os rótulos como parâmetros de comparação de áreas idênticas devido a impossibilidade de realizar uma comparação documento-documento, já que os trabalhos não incluem a lista de documentos das respectivas coleções.

Assim sendo, na base SIGIR, das 29 áreas presentes no total (ao longo de todo o período de 25 anos), a proposta encontrou 28 áreas distintas. A área não encontrada em nenhuma edição foi “Sistemas Gerenciadores de Bancos de dados & Recuperação da Informação”. No trabalho fonte, essa área possuía cinco documentos no total, distribuídos em cinco edições diferentes, ou seja, um documento em cada uma das cinco. Este pode ter sido o motivo da não ocorrência do tema. Outra possiblidade é o rótulo dado, que possui interseções com outros como “Bancos de dados” e “Recuperação da Informação conceitual”.

Considerando o número de áreas presentes ano a ano (incluindo áreas repetidas em anos diferentes), foram encontradas 317 áreas de 353 presentes no artigo fonte. Vale relembrar que a metodologia dos autores consistia de separar a coleção completa em áreas e somente depois dividir os documentos por edição. Essa divisão fez com que áreas presentes em diversos anos contivessem apenas um ou dois documentos. A abordagem da proposta, ao contrário, divide cada edição em tópicos.

Na base SBBD, a fonte considerou todas as áreas distintas, mesmo as que continham transições. Baseado nisso, o número de áreas existentes ano a ano (incluindo aparições da mesma área em anos distintos) foi de 84 no artigo fonte e a proposta acabou encontrando 88 áreas. Apesar de encontrar um número maior de grupos, duas áreas equivalentes não foram encontradas pela proposta: A área 6 de 2008, que não possui um rótulo definido pelos autores, sendo chamada de “Tema 6” e a área 1 de 2009, “Consultas utilizando inferência probabilística”. Não é possível inferir a quantidade de documentos contidos nessas áreas pelos trabalhos fonte. Uma possibilidade é a de ter havido uma fusão com outra área, como “Motores de busca” que também pode abranger consultas. As áreas adicionais encontradas podem ser fruto de divisões internas nos temas originais.

#### *6.5.5.3 Análise temporal*

A Figura 22 e Figura 23 já mostram uma visualização da evolução temporal obtida pela proposta e das evoluções inferidas originalmente.

Para a base SIGIR, a proposta conseguiu identificar 232 transições evolucionárias das 317 transições existentes no trabalho fonte. Além disso foi possível encontrar relações não presentes na fonte (por exemplo, “Bancos de dados” entre 1999 e 2002). Este resultado pode ser oriundo da força de transição adotada. Das transições encontradas tanto na fonte quanto na proposta a menor força de transição foi 0,615 enquanto a maior foi de 0,923. Há que se levar em conta que na presente proposta considera-se as relações entre áreas com pesos, enquanto no trabalho fonte apenas transições rígidas de áreas idênticas ao longo do tempo foram consideradas (ramificações ou evolução de temas não são exibidos).

Para a base SBBD, a proposta identificou 76 transições evolucionárias das 21 transições existentes na fonte, sendo 15 ligando os mesmos grupos nos dois casos. Da mesma forma que a base anterior também é possível visualizar algumas novas ligações não existentes originalmente. Das transições presentes nos dois casos a menor força de transição encontrada foi de 0,405 enquanto a maior foi de 0,856. Entre as transições presentes apenas na proposta a maior força foi de 0,856. O fato de haver mais transições do que originalmente pode ter diversas causas. Primeiro, os autores evidenciaram as transições entre temas distintos visto que temas com rótulos claramente semelhantes não exibiam arestas entre si. Depois, a escolha do peso faz aparecer arestas mais “fracas” que de outro modo não apareceriam, mas que são importantes para a visualização de ramificações e relações multidisciplinares entre os tópicos. Por último, a fusão de alguns temas pelos autores pode ter diminuído o número de áreas (alguns rótulos exibem mais de um tema, por exemplo) ou a proposta pode ter dividido temas que para os especialistas são muito relacionados.

Levando em conta os dados dos trabalhos fonte como conjunto verdade, segure-se então aumentar a restrição da força de transição para valores maiores, indicando áreas temporalmente mais acopladas para acompanhar determinados temas. O uso de pesos menores pode ser usado para uma análise multidisciplinar ou para conhecer melhor as relações existentes ao longo do tempo, como mudanças de paradigma ou ramificações.

# 7 Trabalhos Correlatos

Alguns estudos já começaram a vislumbrar a importância da adoção de mais de um tipo de fonte de dados na análise do desenvolvimento científico-tecnológico. Como suporte ao uso de fontes diversificadas vale mencionar o trabalho de (CALLAERT; LOOY; VERBEEK, 2006), que examina a relação da ciência com a tecnologia através das citações encontradas em patentes. Como resultado descobre que a maioria das citações que não referenciam outras patentes fazem referência a artigos de periódicos e conferências. No trabalho de Callaert et al (2006) também se examinam referências entre patentes e artigos de periódicos para identificação da relação entre as patentes e os domínios científicos. Foram criados dois indicadores para quantificar o número de domínios científicos com os quais uma patente interage e o número de tecnologias com as quais um campo científico interage. No entanto, além de se limitar somente às citações de artigos das patentes, se utiliza de um nível alto de abstração de classificação para a definição de domínios. Os resultados apontam que os domínios de biotecnologia, farmacêutica e tecnologia da informação são as áreas tecnológicas mais relacionadas com a ciência. Por fim, no trabalho de Bhattacharya et al. (2003) são realizadas análises separadamente em artigos e patentes para estabelecer a relação existente entre eles (BHATTACHARYA; KRETSCHMER; MEYER, 2003). São utilizados dois tópicos de pesquisas que resultaram em criações de inovação tecnológica para captura dos documentos relacionados e se chegou à conclusão de que a literatura científica se foca mais em técnicas enquanto que as patentes têm foco maior em aplicação. Todos estes trabalhos já começam a integrar fontes de ciência e de tecnologia em suas análises mesmo que somente por referência e de maneira limitada.

Muitas abordagens para identificação de tópicos e de desenvolvimento são baseadas na criação de redes de referências entre documentos e agrupamento para encontrar similaridade entre eles. Alguns tipos de redes utilizadas são:

- Redes de citação: Rede de documentos onde, dados dois nós A e B, uma aresta entre eles é criada se existe uma citação do documento B no documento A ou vice-versa.
- Redes de co-citação: Rede de documentos onde, dados dois nós A e B, uma aresta entre eles é criada se um terceiro nó C contém citações de A e B.

- Redes de acoplamento bibliográfico: Rede de documentos onde, dados dois nós D e E, uma aresta entre eles é criada sempre que D e E contiverem uma citação a um nó C em comum.
- Redes de coautoria: Rede de pesquisadores, onde dados dois nós A e B, uma aresta entre eles é criada sempre que houver uma publicação científica onde ambos estão listados como autores.

A Figura 6 ilustra os tipos de ligações das redes de citação, co-citação e de acoplamento bibliográfico conforme a explicação anterior.

A etapa de agrupamento depende então do tipo de rede utilizada e o tipo de agrupamento resultante dependerá do significado implícito das ligações de cada grafo montado. A identificação dos tópicos é realizada então via exame dos grupos resultantes do processo de agrupamento.

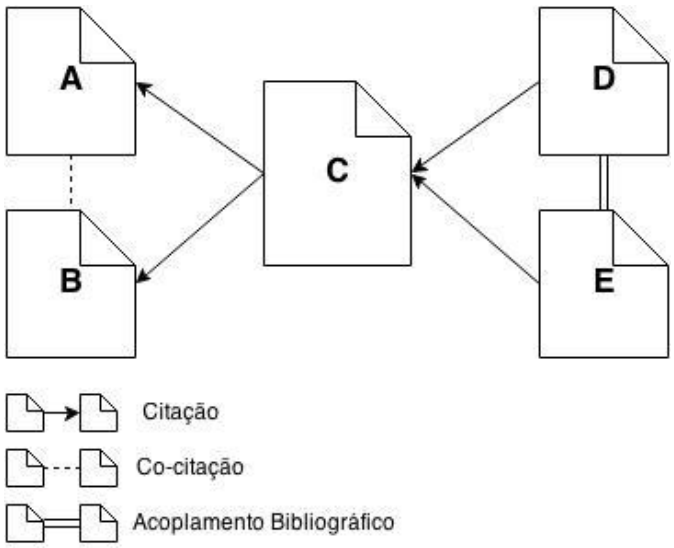



Figura 25: Tipos de citações decorrentes da interação entre documentos. Fonte: Adaptado de (SHIBATA; KAJIKAWA, 2009)

Como exemplo de trabalhos utilizando este processo temos o de Courtial et al. (1993), que faz uso de redes agrupadas de patentes para encontrar grupos centrais (com grande número de ligações) e densos (com grande número de ligações em relação a todas as possíveis) na rede, o que se presume corresponda a áreas mais importantes (COURTIAL et al., 1993). Este é um trabalho que possui como limitações levar em conta apenas o âmbito da

inovação, utilizar apenas títulos de patentes como medida de similaridade e necessitar de identificação manual de áreas. Tseng et al. (2009) fazem um estudo comparativo de métodos de detecção de temas “quentes” na ciência. Através do uso de redes de grupos de co-citação em artigos e da identificação manual dos tópicos. São contrastados índices que indicam se um tema é tendência com a opinião de especialistas na área. Como resultados obtiveram, além do índice mais consistente baseado em regressão linear, que a origem dos artigos (país de origem, tipo de periódico) não teve influência no resultado das avaliações. Ressalta também que a criação de formas automáticas de se monitorar o desenvolvimento dos tópicos, especialmente os mais importantes, seria de grande ajuda (TSENG et al., 2009). Já Shibata et al. (2009) realizam um estudo comparativo dos diferentes tipos de redes referências para avaliar qual seria o melhor para identificar frentes de pesquisa científica, ou seja, pesquisas com aumento de interesse pela comunidade científica. As redes comparadas foram a rede de citação, a de co-citação e a de acoplamento bibliográfico. Para o objetivo de encontrar pesquisas emergentes a que se provou melhor foi a de citação. Há intervenção manual na identificação dos subdomínios pertencentes aos domínios utilizados como escopo do estudo o que torna necessário o conhecimento das áreas mais detalhadamente por parte dos autores (SHIBATA et al., 2009).

Alguns outros trabalhos focam em visualizações como meio de suporte a identificação e acompanhamento do desenvolvimento por parte de especialistas. Essa abordagem é semiautomática visto que todas as inferências têm de passar pelo crivo pessoal de alguém com conhecimentos específicos no domínio.

Exemplos de trabalho nessa linha incluem as pesquisas de (CHEN, 2005), que cria uma rede de grupos de co-citação, desta vez limitada por artigos e desenvolve uma visualização desta rede onde o usuário deve identificar por si as áreas e tirar suas próprias conclusões a partir dos dados apresentados. Também admite que os usuários precisam de ferramentas que transformem a vasta quantidade de dados em mensagens limpas e instrutivas e indica que a ligação com invenções seria de valor central para a política em C&T e para a área de difusão de conhecimento. Kurosawa & Takama (2011) propõem um sistema de visualização de redes de coautoria científica que facilitem a identificação de pesquisadores emergentes e de supervisores de pesquisa (KUROSAWA; TAKAMA, 2011). Neste trabalho a identificação das áreas e do desenvolvimento dos pesquisadores fica a cargo do usuário, mas são criadas

funções como a de encontrar pesquisadores similares para dar suporte a estas tarefas. Morris et al. (2003) desenvolvem uma ferramenta de visualização de áreas de pesquisas científicas emergentes em séries temporais. Os documentos são agrupados por quantidade de palavras em comum (MORRIS et al., 2003). A identificação das áreas neste caso é realizada manualmente através de exame nos títulos dos artigos.

Outra tarefa que tem um relacionamento muito próximo com a identificação de áreas e de seu desenvolvimento é a de análise de crescimento de frentes de pesquisa e de tecnologias emergentes. As frentes de pesquisa são caracterizadas por novos tópicos ainda em desenvolvimento criados a partir de outros já em estado de maturidade. As tecnologias emergentes compreendem as inovações com crescimento em registros e demanda de mercado.

Trabalhos nesta área compreendem pesquisas como a de (DAIM et al., 2006) que utiliza diferentes metodologias em conjunto para detectar tecnologias emergentes utilizando patentes como material. Mais uma vez acontece uma abordagem semiautomática onde o autor mistura estatísticas, identificação manual de áreas e análises de mercado para descobrir as tecnologias em recente crescimento. Como se trata de uma análise comparativa, como resultados temos uma lista de combinações de métodos que podem ser usadas em diferentes cenários, necessitando de um especialista para a escolha. (SMALL, 2006) explora a possibilidade do uso de redes de grupos de co-citação através de fatias de tempo para rastrear o crescimento de áreas na ciência. Os resultados indicam que há valor limitado na estratégia e novamente é necessária intervenção manual para a identificação das áreas ao longo do tempo, assim como seus desmembramentos. Shibata et al. (2008) criam métricas para auxílio na detecção de frentes de pesquisa emergentes baseadas em medidas topológicas em redes de citação de publicações científicas. Identifica dois tipos de inovação: Incremental, onde a área cresce em tamanho; e por ramificação, onde a área se desmembra em duas ou mais. Reconhece também que a identificação de domínios por visualização requer julgamento implícito dos usuários e que existe aumento na demanda por inteligência capaz de descobrir domínios de pesquisa emergentes e seus respectivos tópicos. Conclui afirmando que o uso de fontes de dados adicionais como os de inovação ajudaria a entender mais profundamente o progresso técnico (SHIBATA et al., 2008). Bengisu & Nekhili (2006) já utilizam bases de patentes e de artigos em sua pesquisa para identificar e prever tecnologias emergentes

(BENGISU; NEKHILI, 2006). Neste estudo a identificação dos tópicos é dada pelas palavras-chave presentes na base de dados utilizada, o que não costuma ocorrer em todas as fontes de dados de C&T atuais, além da possibilidade de não descrever corretamente o conteúdo do documento (COURTIAL; CALLON; SIGOGNEAU, 1992). Tanto a parte de acompanhamento do desenvolvimento quanto a previsão são realizados de maneira manual de acordo com a tecnologia.

Já estão sendo criadas até novas formas de se analisar dados de C&T utilizando o tópico como matéria-prima como em (MANN; MIMNO; MCCALLUM, 2006) que tem como objetivo estender a bibliometria dos periódicos para os tópicos, criando novas medidas de influência a nível de área através de novas métricas. O fator de impacto do tópico, a difusão do tópico e transferência do tópico são alguns exemplos de métricas existentes em periódicos estendidas para uso a nível de domínios.

Finalmente, estudos de identificação e rastreamento de tópicos de interesse utilizando técnicas de mineração de texto em cima de um único tipo de base podem ser encontrados em (KAUER, 2013), que utiliza uma técnica de mineração de textos temporal para rastrear o desenvolvimento dos tópicos de interesse do Simpósio Brasileiro de Banco de Dados (SBBD), assim como suas ramificações e junções ao longo do tempo. A técnica utilizada é proveniente do trabalho de Mei & Zhai (2005), que também a utiliza para analisar a distribuição de temas ao longo do tempo dos artigos da conferência *Knowledge Discovery and Data Mining* (KDD) e dos tópicos relacionados ao tsunami da Ásia em 2005 (MEI; ZHAI, 2005). A técnica utilizada por ambos tem a limitação de ter que se definir a priori o número de temas para agrupamento dos documentos e de ter que definir alguns parâmetros da função empiricamente, exigindo uma intervenção humana de acordo com a coleção de documentos utilizada. As coleções utilizadas também devem ser homogêneas, ou seja, ter a mesma estrutura e/ou serem provenientes da mesma base.

Assim, a maioria dos trabalhos que tentam realizar uma identificação de áreas de pesquisa utiliza meios semiautomáticos. Esses meios consistem na utilização de uma ferramenta (agrupamentos, visualizações, mineração) que automatiza processos de agrupamento mas que necessitam de posterior conhecimento especializado para interpretação dos resultados. Por exemplo, em grupos de co-citação ainda é necessário que o pesquisador tenha um conhecimento dos documentos e autores da área para que possa

defini-la. Vários conhecimentos por parte dos usuários ainda são necessários, como por exemplo a definição da quantidade de áreas presentes na coleção, interpretação dos tópicos, definição de parâmetros que dependem de informação do domínio. Dificilmente um pesquisador que esteja estudando um novo domínio ou que não seja um especialista na área saberá definir quantas áreas estão presentes em uma coleção ou identificar tópicos por autor. Nenhum dos trabalhos apresenta uma abordagem totalmente automática, que possa ser utilizada por leigos na área ou que não exijam análise posterior ou conhecimento prévio.

Outra questão é o uso de dados bibliométricos como as citações. Esses dados nem sempre estão disponíveis ou completos. Supõe-se que os autores citam as mesmas fontes, o que seria uma indicação da área, ou que a base de dados possua todos os documentos de uma área formando uma rede completa de ligações. Na realidade, muitas vezes um autor não referencia todos os trabalhos da área e na maioria dos casos uma base ou coleção de documentos conterá apenas um subconjunto do universo de documentos pertencentes a determinada área.

Desta maneira, o presente trabalho visa a construção de um método verdadeiramente automático na identificação das áreas de pesquisa desde o agrupamento de dados inicial até a posterior interpretação dos resultados e análise temporal. Incluindo também a independência de fonte, ou seja, não é necessária uma rede completa de documentos, conhecimentos do domínio ou de parâmetros atrelados a coleção e ao escopo. Como todo mecanismo automático, deve sempre dispensar conhecimento especializado para a sua execução.

# 8 Conclusão

Este trabalho objetivou a identificação automática de áreas científicas para a gestão estratégica desde o nível do pesquisador até aos investidores e do próprio País. O suporte que a proposta fornece pode ser utilizado tanto no âmbito pessoal, para organização e exploração de coleções de documentos por tema, até ao uso em grandes bases de dados, para exploração e detecção dos melhores investimentos. A solução também ajuda ao rastreio de áreas novas de pesquisa que ainda não estão bem indexadas e atualmente, com o crescimento do uso de dados não-estruturados e semi-estruturados, no acompanhamento e exploração de pesquisas que se utilizam de meios de divulgação mais rápidos.

Foram estudadas e avaliadas técnicas para detecção de tópicos de pesquisa em meio a coleções textuais científico-tecnológicas, identificação do número de áreas presentes em uma coleção e rotulagem de grupos de documentos para facilitar a compreensão sobre o tema contido. Dentre cada uma destas áreas, foram selecionadas as melhores técnicas e adaptadas ao ambiente científico e as necessidades da proposta e objetivo. Quando não existente, foram criadas novas técnicas parar servir ao propósito final, sempre priorizando a mínima intervenção humana nos resultados. Outras prioridades e fatores utilizados para atingir maior robustez foram o uso de amostragem sempre que possível, poupando recursos computacionais (escalabilidade); Independência de linguagem e fonte de dados, todo o processo pode ser usado com vários idiomas e dados pois não necessita de processamentos linguísticos específicos (por exemplo, árvores sintáticas); e Modularidade, pois todas as fases dos processos envolvidos podem ser modificadas sem prejuízo dos dados que já existem, o que facilita a utilização em ambientes dinâmicos.

Avaliações qualitativas foram realizadas para obtenção de evidências que mostrassem a possiblidade de identificação e representação de áreas de pesquisa presentes em coleções textuais de forma não-supervisionada e satisfatória aos usuários. Estas mesmas avaliações indicam que é possível a exploração de coleções através da detecção automática de áreas de pesquisa, incluindo formas de rotulagem que ajudam os usuários a entenderem melhor o conteúdo de cada área com termos específicos.

Finalmente, a combinação de diversas técnicas, novas e existentes, abriu um caminho para a resolução de um problema antigo como a análise temática de dados. Os resultados

alcançados indicam um avanço em comparação com o paradigma vigente, ao mesmo tempo que alcança uma maior flexibilidade e modularidade em relação às soluções existentes, manuais ou semiautomáticas.

## 8.1 Trabalhos Futuros

Num mundo com cada vez mais dados e bancos de dados sem estrutura rígida, seria bom que houvessem mais aplicações que tornassem acessíveis a todos as inovações construídas através de melhores interfaces, possibilitando uma navegação exploratória sobre os dados, por exemplo.

Oportunidades presentes envolvem primeiramente a integração com outras formas de agrupamento por modelagem de tópicos, como algoritmos que funcionam em ambientes distribuídos ou dinâmicos, ou seja, com modificações em tempo real.

Outra possiblidade é o uso de processos hierárquicos para construção de árvores temáticas evitando assim a seleção do número de áreas, mas que por outro lado pode ofuscar a visualização dos resultados.

O estudo da relação entre áreas também pode ser ampliado com a extrapolação e métricas que avaliem o crescimento ou declínio futuro de áreas de pesquisa baseando-se nas tendências temporais. Estudos de prospecção e cooperação também podem utilizar as ramificações, migrações e fusões do grafo temporal para analisar o comportamento das áreas e da comunidade.

A rotulagem também pode ser enriquecida através da exploração da semântica entre termos e do uso de ontologias para estabelecer relações entre conceitos abrindo possibilidades de sumarização e estudos em multidisciplinariedade. Assim, a semântica extraída de bases externas poderia ser um fator na escolha dos rótulos mais representativos, associando áreas distintas ou identificando e ajudando o usuário no entendimento da terminologia da área.

Por fim, a aplicação em outros tipos de dados pode ser usada para mostrar outras facetas do presente trabalho. Aqui se trabalhou em um ambiente acadêmico onde existe um certo rigor no texto e na terminologia utilizada. O uso com textos coloquiais e informais ou em mensagens curtas pode abrir novas possiblidades, como por exemplo a identificação de eventos em *tweets* realizada por (LAUAND, 2016) entre outras nas áreas de redes sociais, altmetria, acervos pessoais e bases relacionadas à saúde. Exemplos de aplicações nessas

seriam, por exemplo, a evolução dos assuntos nas redes sociais por tópico, identificação de áreas de pesquisa na rede para melhorar os dados de altmetria, a exploração dos *e-mails* pessoais e o acompanhamento na evolução da elaboração de novas vacinas e tratamentos.

## 8.2 Limitações

Neste trabalho não foi possível realizar testes e avaliações em todos os domínios científicos ou fontes. A validade dos resultados, apesar de sugerida, não pôde ser então atestada para qualquer campo da ciência. Porém todas as áreas possuem uma certa terminologia distinta, possíveis diferenças poderiam ser encontradas num domínio de humanas por exemplo, com certas árteas mais semelhantes umas as outras. Apesar do uso de diferentes fontes, estas também não foram esgotadas. Atualmente cresce o número de publicações científicas em meios alternativos, como *blogs* e mídias sociais que não foram utilizadas aqui.

Em relação as técnicas utilizadas, pode-se ressaltar como limitações: (i) o tempo de processamento, que impede que a identificação das áreas seja feita *online* (isso é verdade para ambientes pessoais, não foram realizados testes em ambientes corporativos, com servidores por exemplo); (ii) O intervalo do número de áreas que é testado na seleção do número de áreas, que na teoria pode incluir todas as possibilidades mas na prática deve se limitar a um certo intervalo para agilizar o processo principalmente para grandes volumes de dados ou máquinas com processamento inferior. Com relação à rotulagem, o uso do modelo *bag-of-words* que trata todas as palavras de forma independente pode ser considerado uma limitação pois não leva em consideração a semântica ou contexto de cada termo.

# Referências


ABRIL, A. **Enciclopédia de Atualidades**. [s.l.] São Paulo: Abril, 2012.

AGRAWAL, R. et al. Automatic subspace clustering of high dimensional data for data mining applications. 1998.

ANKERST, M. et al. OPTICS: ordering points to identify the clustering structure. **ACM Sigmod Record**, 1999.

ARABIE, P.; HUBERT, L. AN OVERVIEW OF COMBINATORIAL DATA. **Clustering and classification**, 1996.

BAE, E.; BAILEY, J. Coala: A novel approach for the extraction of an alternate clustering of high quality and high dissimilarity. **Conference on Data Mining (ICDM'06)**, 2006.

BALL, G.; HALL, D. ISODATA, a novel method of data analysis and pattern classification. 1965.

BELLMAN, R. **Dynamic programming**. [s.l.] Dover Publications, 2003.

BEN-DAVID, S.; PÁL, D.; SIMON, H. Stability of k-means clustering. **International Conference on**, 2007.

BENGISU, M.; NEKHILI, R. Forecasting emerging technologies with the aid of science and technology databases. **Technological Forecasting and Social Change**, v. 73, n. 7, p. 835–844, set. 2006.

BERRY, M. W. J. K. **Text Mining Applications and Theory**. West Sussex, UK: John Wiley & Sons, 2010.

BERTONI, A.; VALENTINI, G. Random projections for assessing gene expression cluster stability. **Proceedings. 2005 IEEE International**, 2005.

BHATTACHARYA, S.; KRETSCHMER, H.; MEYER, M. Characterizing intellectual spaces between science and technology. **Scientometrics**, v. 58, n. 2, p. 369–390, 2003.

BLEI, D. Probabilistic topic models. **Communications of the ACM**, 2012.

BLEI, D.; LAFFERTY, J. Dynamic topic models. **23rd international conference on Machine learning**, 2006.

BLEI, D.; LAFFERTY, J. Topic models. **: classification, clustering, and applications**, 2009.

BLEI, D. M.; NG, A Y.; JORDAN, M. I. Latent Dirichlet Allocation. **Journal of Machine Learning Research**, v. 3, p. 993–1022, 2003.

BRUNET, J.; TAMAYO, P.; GOLUB, T. Metagenes and molecular pattern discovery using matrix factorization. **Proceedings of the**, 2004.

CALLAERT, J.; LOOY, B. VAN; VERBEEK, A. Traces of prior art: An analysis of non-patent references found in patent documents. **Scientometrics**, v. 69, n. 1, p. 3–20, 2006.

CASELLA, G.; GEORGE, E. Explaining the Gibbs sampler. **The American Statistician**, 1992.

CHANG, J. et al. Reading Tea Leaves: How Humans Interpret Topic Models. **Advances in Neural Information Processing Systems 22**, p. 288--296, 2009.

CHEN, C. CiteSpace II : Detecting and Visualizing Emerging Trends. **Journal of the American Society for Information Science and Technology**, v. 57, n. 3, p. 359–377, 2006a.

CHEN, C. CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. **Journal of the American Society for Information Science and Technology**, v. 57, n. 3, p. 359–377, 1 fev. 2006b.

COLE, R. Clustering with genetic algorithms. 1998.

COURTIAL, J. P. et al. The use of patent titles for identifying the topics of invention and forecasting trends. **Scientometrics**, v. 26, n. 2, p. 231–242, 1993.

DAIM, T. U. et al. Forecasting emerging technologies: Use of bibliometrics and patent analysis. **Technological Forecasting and Social Change**, v. 73, n. 8, p. 981–1012, out. 2006.

DOWNEY, D. et al. Active Learning with Constrained Topic Model. **Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces**, p. 30–33, 2014.

ESTER, M. et al. Incremental clustering for mining in a data warehousing environment. **VLDB**, 1998.

FAGIN, R.; KUMAR, R.; SIVAKUMAR, D. Comparing top k lists. **SIAM Journal on Discrete Mathematics**, 2003.

GREENE, D.; O'CALLAGHAN, D.; CUNNINGHAM, P. How Many Topics? Stability Analysis for Topic Models. **Machine Learning and Knowledge Discovery in Databases**, 2014.

GUHA, S.; RASTOGI, R.; SHIM, K. CURE: an efficient clustering algorithm for large databases. **ACM SIGMOD Record**, 1998.

HAN, J.; PEI, J.; KAMBER, M. Data mining: concepts and techniques. 2011.

HOFMANN, T. Probabilistic latent semantic indexing. **SIGIR '99: Proceedings of the 22nd annual international conference on research and development in information**

**retrieval**, p. 50–57, 1999.

JACCARD, P. The distribution of the flora in the alpine zone. **New phytologist**, 1912.

JANSSENS, F.; GLÄNZEL, W.; MOOR, B. DE. Dynamic hybrid clustering of bioinformatics by incorporating text mining and citation analysis. **discovery and data mining**, 2007.

KAUER, V. A. Evolução dos Temas de Interesse do SBBD ao Longo dos Anos. **Simpósio Brasileiro de Banco de Dados**, p. 1–6, 2013.

KONTOSTATHIS, A. et al. A survey of emerging trend detection in textual data mining. **Survey of text**, 2004.

KUHN, H. The Hungarian method for the assignment problem. **Naval research logistics quarterly**, 1955.

KUROSAWA, T.; TAKAMA, Y. Predicting Researchers' Future Activities Using Visualization System for Co-authorship Networks. **2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology**, p. 332–339, ago. 2011.

LANDAUER, T.; DUMAIS, S. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. **Psychological review**, 1997.

LANGE, T. et al. Stability-based validation of clustering solutions. **Neural computation**, 2004.

LAU, J. H. et al. Best Topic Word Selection for Topic Labelling. **Proceedings of the 23rd International Conference on Computational Linguistics: Posters**, n. August, p. 605–613, 2010.

LAU, J. H. et al. Automatic labeling of topic models. **Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics**, p. 1536–1545, 2011.

LAUAND, B. **Contextualização das Informações das Mídias Sociais para Uso em Situações de Emergência**. [s.l.] Universidade Federal do Rio de Janeiro, 2016.

LAW, M. H.; JAIN, A. K. Cluster validity by bootstrapping partitions. **Tech. Rep. MSUCSE-03-5.**, 2003.

LEVINE, E.; DOMANY, E. Resampling method for unsupervised estimation of cluster validity. **Neural computation**, 2001.

LLOYD, S. Least squares quantization in PCM. **IEEE transactions on information theory**, 1982.

LUHN, H. P. A Statistical Approach to Mechanized Encoding and Searching of Literary

Information. **IBM Journal of Research and Development**, v. 1, n. 4, p. 309–317, out. 1957.

MACQUEEN, J. Some methods for classification and analysis of multivariate observations. **Proceedings of the fifth Berkeley symposium on**, 1967.

MANN, G. S.; MIMNO, D.; MCCALLUM, A. Bibliometric impact measures leveraging topic analysis. **Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries - JCDL '06**, p. 65, 2006.

MANNING, C. D. et al. **Introduction to Information Retrieval**. New York, New York, USA: Cambridge University Press, 2007.

MANYIKA, J. et al. Big data: The next frontier for innovation, competition, and productivity. 2011.

MEI, Q. et al. A probabilistic approach to spatiotemporal theme pattern mining on weblogs. **Proceedings of the 15th international conference on World Wide Web - WWW '06**, p. 533, 2006.

MEI, Q.; SHEN, X.; ZHAI, C. Automatic labeling of multinomial topic models. **Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '07**, p. 490, 2007.

MEI, Q.; ZHAI, C. Discovering evolutionary theme patterns from text: an exploration of temporal text mining. **… conference on Knowledge discovery in data mining**, p. 198–207, 2005.

MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E INOVAÇÃO, M. **Expansão e consolidação do sistema nacional de C,T&I**. Disponível em: <http://www.mct.gov.br/index.php/content/view/73410/I_Expansao_e_Consolidacao_do_Sistema_Nacional_de_CT_I.html>. Acesso em: 21 maio. 2016.

MORRIS, S. A. et al. Timeline Visualization of Research Fronts. **Journal of the American Society of Information Science Technology**, v. 54, n. 5, p. 413–422, 2003.

NEWMAN, D. et al. Automatic evaluation of topic coherence. **Language Technologies: The …**, 2010.

OLIVEIRA, J. DE; PEDRYCZ, W. Advances in fuzzy clustering and its applications. 2007.

PRIYA, M. B., KUMARAVEL, A. Methodologies for Trend Detection Based on Temporal Text Mining. v. 2, n. 4, p. 540–554, 2013.

PRIYA, M.; KUMARAVEL, A. Methodologies for Trend Detection Based on Temporal

Text Mining. 2013.

R. NG, J. H. Efficient and Effective Clustering Algorithms for Spatial Data Mining. **VLDB Conference**, 1994.

RAJARAMAN, A.; ULLMAN, J. D. Data Mining. In: **Mining of Massive Datasets**. Cambridge: Cambridge University Press, 2011. p. 1–17.

RAMAGE, D.; MANNING, C. D.; DUMAIS, S. Partially labeled topic models for interpretable text mining. **Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining**, p. 457–465, 2011.

REDE NACIONAL DE ENSINO E PESQUISA. **Mercadante destaca avanços e diz que deixa o MCTI em boas mãos**. Disponível em: <http://portal.rnp.br/web/rnp/noticias/-/rutelistaconteudo/Mercadante-destaca-avancos-e-diz-que-deixa-o-MCTI-em-boas-maos/551716_o80B;jsessionid=2BBE0CCABA1E0527BA28E9FB6000ED3F.inst2>. Acesso em: 20 abr. 2016.

SHIBATA, N. et al. Detecting emerging research fronts based on topological measures in citation networks of scientific publications. **Technovation**, v. 28, n. 11, p. 758–775, nov. 2008.

SHIBATA, N.; KAJIKAWA, Y. Comparative study on methods of detecting research fronts using different types of citation. **Journal of the …**, v. 60, n. 1971, p. 571–580, 2009.

SMALL, H. Tracking and predicting growth areas in science. **Scientometrics**, v. 68, n. 3, p. 595–610, 2006.

SMEATON, A. et al. Analysis of papers from twenty-five years of sigir conferences: what have we been doing for the last quarter of a century? **ACM SIGIR Forum**, 2002.

SOARES. **Ciência Aberta**. Disponível em: <<http://cienciahoje.uol.com.br/blogues/bussola/2012/03/ciencia-aberta>. Acesso em: 20 abr. 2016.

SOLINGEN, R. VAN; BERGHOUT, E. The Goal/Question/Metric Method: A Practical Guide for Quality Improvement of Software Development. 1999.

SORENSEN, T. {A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on. **Biol. Skr.**, 1948.

SPARCK JONES, K. A STATISTICAL INTERPRETATION OF TERM SPECIFICITY AND ITS

APPLICATION IN RETRIEVAL. **Journal of Documentation**, v. 28, n. 1, p. 11–21, jan. 1972.

STEYVERS, M.; GRIFFITHS, T. Probabilistic topic models. **Handbook of latent semantic analysis**, 2007.

TANG, J. et al. Arnetminer: An expertise oriented search system for web community. **Proceedings of the 2007**, 2007.

THE NATIONAL ACADEMIES PRESS. **President Barack Obama's Speech to the National Academy of Sciences**. Disponível em: <<http://notes.nap.edu/2013/04/30/president-barack-obamas-speech-to-the-national-academy-of-sciences-full-transcript/#.U3_wdx8Q6zU>. Acesso em: 21 maio. 2016.

TSENG, Y.-H. et al. A comparison of methods for detecting hot topics. **Scientometrics**, v. 81, n. 1, p. 73–90, 18 mar. 2009.

WAAL, A. DE; BARNARD, E. Evaluating topic models with stability. **19th Annual Symposium of the Pattern**, 2008.

WANG, Q. et al. Group matrix factorization for scalable topic modeling. **Proceedings of the 35th international ACM**, 2012.

WORLD BANK. **WORLD BANK**. Disponível em: <http://www.worldbank.org/>. Acesso em: 21 maio. 2016.

WORLD INTELLECTUAL PROPERTY ORGANIZATION - WIPO. **World Intellectual Property Indicators**. Disponível em: <http://www.wipo.int/export/sites/www/freepublications/en/intproperty/941/wipo_pub_941_2013.pdf>. Acesso em: 1 maio. 2014.

ZHANG, T.; RAMAKRISHNAN, R.; LIVNY, M. BIRCH: an efficient data clustering method for very large databases. **ACM Sigmod Record**, 1996.

# Apêndices

## APÊNDICE A – MODELO DO QUESTIONÁRIO DE AVALIAÇÃO

### PESQUISA DE AVALIAÇÃO

Nome:

### Instruções

De acordo com o os documentos e termos apresentados, para cada tópico escolha um label que, na sua opinião, melhor represente o conceito contido nele. Em seguida serão apresentadas algumas sugestões de labels onde devem ser associadas notas de 1 a 5 em duas escalas, conforme a seguinte legenda:

- Escala 1 – O quão bem o label representa o tópico correspondente.
  - 5 = Excelente, representa exatamente o assunto do tópico
  - 4 = Muito bom, representa bem o assunto do tópico
  - 3 = Bom, É possível saber o assunto do tópico através dele
  - 2 = Não muito bom, Relacionado mas ruim como label
  - 1 = Ruim, completamente inapropriado
- Escala 2 – O quanto o label sugerido se assemelha com o label que você escolheu
  - 5 = Os labels são quase ou exatamente idênticos
  - 4 = Os labels são muito parecidos em sentido
  - 3 = Os labels são relacionados
  - 2 = Tem alguma relação, mas o sentido é diferente
  - 1 = Não se assemelham em nada

Para cada item identificado abaixo, faça um círculo ao redor do número à direita que melhor combina com seu julgamento de qualidade.

Tópico 0: palavra1 palavra2 palavra3

Label manual:

| Label(s) | Escala 1 (Representação) | | | | | Escala 2 (Semelhança) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Ruim | | Bom | | Excelente | Pouco | | Mediano | | Muito |
| Descrição do item | 1 | 2 | 3 | 4 | 5 | 1 | 2 | 3 | 4 | 5 |
| Descrição do item | 1 | 2 | 3 | 4 | 5 | 1 | 2 | 3 | 4 | 5 |
| Descrição do item | 1 | 2 | 3 | 4 | 5 | 1 | 2 | 3 | 4 | 5 |
| Descrição do item | 1 | 2 | 3 | 4 | 5 | 1 | 2 | 3 | 4 | 5 |
| Descrição do item | 1 | 2 | 3 | 4 | 5 | 1 | 2 | 3 | 4 | 5 |
| Descrição do item | 1 | 2 | 3 | 4 | 5 | 1 | 2 | 3 | 4 | 5 |
| Descrição do item | 1 | 2 | 3 | 4 | 5 | 1 | 2 | 3 | 4 | 5 |
| Descrição do item | 1 | 2 | 3 | 4 | 5 | 1 | 2 | 3 | 4 | 5 |

## APÊNDICE B – MODELO DO QUESTIONÁRIO COM DADOS REAIS

### PESQUISA DE AVALIAÇÃO

Nome:

### Instruções

De acordo com o os documentos e termos apresentados, para cada tópico escolha um label que, na sua opinião, melhor represente o conceito contido nele. Em seguida serão apresentadas algumas sugestões de labels onde devem ser associadas notas de 1 a 5 em duas escalas, conforme a seguinte legenda:

- Escala 1 – O quão bem o label representa o tópico correspondente.
  - 5 = Excelente, representa exatamente o assunto do tópico
  - 4 = Muito bom, representa bem o assunto do tópico
  - 3 = Bom, È possível saber o assunto do tópico através dele
  - 2 = Não muito bom, Relacionado mas ruim como label
  - 1 = Ruim, completamente inapropriado
- Escala 2 – O quanto o label sugerido se assemelha com o label que você escolheu
  - 5 = Os labels são quase ou exatamente idênticos
  - 4 = Os labels são muito parecidos em sentido
  - 3 = Os labels são relacionados
  - 2 = Tem alguma relação, mas o sentido é diferente
  - 1 = Não se assemelham em nada

Para cada item identificado abaixo, faça um círculo ao redor do número à direita que melhor combina com seu julgamento de qualidade.

Tópico 1: graph, graphs, network, nodes, networks, user, large, structure, social, quality

Label manual:

| Label(s) | Escala 1 (Representação) Ruim | | Bom | | Excelente | Escala 2 (Semelhança) Pouco | | Mediano | | Muito |
|---|---|---|---|---|---|---|---|---|---|---|
| Social networks | 1 | 2 | 3 | 4 | 5 | 1 | 2 | 3 | 4 | 5 |
| Social networks, nodes, connection subgraphs | 1 | 2 | 3 | 4 | 5 | 1 | 2 | 3 | 4 | 5 |
| Social networks | 1 | 2 | 3 | 4 | 5 | 1 | 2 | 3 | 4 | 5 |
| Social networks, large social network graphs, nodes | 1 | 2 | 3 | 4 | 5 | 1 | 2 | 3 | 4 | 5 |
| Social networks | 1 | 2 | 3 | 4 | 5 | 1 | 2 | 3 | 4 | 5 |
| Social networks, large social network graphs, nodes | 1 | 2 | 3 | 4 | 5 | 1 | 2 | 3 | 4 | 5 |
| Social networks | 1 | 2 | 3 | 4 | 5 | 1 | 2 | 3 | 4 | 5 |
| Social networks, social network, graph | 1 | 2 | 3 | 4 | 5 | 1 | 2 | 3 | 4 | 5 |
| Social network | 1 | 2 | 3 | 4 | 5 | 1 | 2 | 3 | 4 | 5 |
| social network, large network, data structures | 1 | 2 | 3 | 4 | 5 | 1 | 2 | 3 | 4 | 5 |
| Large network | 1 | 2 | 3 | 4 | 5 | 1 | 2 | 3 | 4 | 5 |

# APÊNDICE C – RESULTADOS KDD

| **Tópico** | identify, disease, medical, identifying, health | optimization, methods, proposed, formulation, show | approach, interesting, show, mining, effective | knowledge, accuracy, detection, sample, available | results, significant, high, probability, problem |
|---|---|---|---|---|---|
| **Rótulos** | | | | | |
| **ET1-tf** | disease progression, disease, disease study | show, dual optimization, propose | text mining, experiments, interesting | work, sample size, intrusion detection | order, results, maximization problem |
| **ET2-tf** | disease, disease progression, modern electronic healthcare records | proposed methods, proposed formulation, proposed dual optimization algorithm | discover interesting evolutionary theme patterns effectively, effectively summarize, experiments show | detection accuracy, work, sample size | order, based models present good results, results |
| **ET3-tf** | disease progression, disease study, disease | proposed methods, proposed formulation, proposed dual optimization | discover interesting evolutionary theme patterns effectively, patterns effectively, theme patterns effectively | detection accuracy, work, sample size | order, based models present good results, results |
| **ET1-mdeg** | disease, disease progression, features | propose, proposed, regression | interesting, text mining, interesting evolutionary | work, performance, spam detection | results, algorithms, order |
| **ET1-medeg** | disease progression, disease, features | propose, optimization algorithm, proposed | text mining, interesting, interesting evolutionary | work, spam detection, performance | results, maximization problem, algorithms |

| | | | | | |
|---|---|---|---|---|---|
| **ET1-deg/tf** | clinical/cognitive measures including mini mental state examination, accurately identified larger disease subsystems implicated, disease assessment scale cognitive subscale | sparse logistic regression embeds feature selection, proposed scheme generally achieves superior results, scale sparse logistic regression logistic regression | discover interesting evolutionary theme patterns effectively, temporal text mining temporal text mining, typically significantly outperforms extant approaches | traditional duplicate detection techniques relying, embedded unbiased sample heuristic, complement expert medical knowledge | based models present good results, influence difference maximization problem, studied influence maximization problem |
| **ET2-deg/tf** | disease state, clinical deterioration, healthcare system | linear discriminant, based methods, optimization algorithm | effectively summarize, interesting themes, text mining | spam detection, background knowledge, small sample | based models, maximization problem, success based |
| **ET3-deg/tf** | clinical/cognitive measures, disease state, clinical deterioration | based regressions, results show, regression/classification results | discover interesting, patterns effectively, interesting evolutionary | detection techniques, duplicate detection, medical knowledge | good results, based models, based predictive |
| **EP1-tf** | classification, alzheimer's disease, neuroimaging | kernel learning, cyber-physical system, classification | clustering, summarization, poset | adversarial classification, spam filtering, data cleaning | information theory, temporal pattern discovery, classification |
| **EP2-tf** | neuroimagining, data mining, alzheimer's disease | classification, data mining, algorithms | summarization, data mining, experimentation | Spam filtering, data mining, concept learning | Pattern discovery, data mining, theory |
| **EP3-tf** | classification, learning, disease | kernel, method, classification | clustering, pattern, summarization | spam, data, learning | clustering, networks, temporal |

| | | | | | |
|---|---|---|---|---|---|
| **EP1-mdeg** | spatial association analysis, classification, disease progression | kernel learning, cyber-physical system, classification | poset, markov random field, gene expression | cost-sensitive learning, naive bayes, sampling | information theory, temporal pattern discovery, classification |
| **EP2-mdeg** | spatial association analysis, classification, disease progression | kernel learning, learning, cyber-physical system | poset, cyber-physical system, markov random field | statistical, cost-sensitive learning, naive bayes | theory, information search and retrieval, information theory |
| **EP1-deg/tf** | niddk liver transplant database, sparse inverse covariance estimation, heterogeneous data source fusion | l_1-norm max-margin markov networks, scalable and sparse kernel method, quadratically constrained quadratic programming | markov random field, minimum description length, tweet entity linking | integer linear programming, disguised missing data, linear classifiers | kulldorff scan statistic, curse of cardinality, sequential pattern mining |
| **EP2-deg/tf** | sparse inverse covariance estimation, niddk liver transplant database, heterogeneous data source fusion | l_1-norm max-margin markov networks, scalable and sparse kernel method, spatial databases and gis | abuse and crime involving computers, information search and retrieval, markov random field | classifier design and evaluation, feature evaluation and selection, integer linear programming | computations on discrete structures, life and medical sciences, information search and retrieval |
| **EP3-deg/tf** | alzheimer's, aided, kernel | nesterov's, and, kernel | poset, model, evolutionary | high-dimensional, selection, learning | k-itemsets, leagues, selection |

| | | | | | |
|---|---|---|---|---|---|
| **Tópico** | system, mining, management, techniques, analysis | rate, set, empirical, study, measure | state, art, approaches, novel, significantly | social, networks, network, users, online | web, search, query, user, queries |
| **Rótulos** | | | | | |
| **ET1-tf** | data mining, detection systems, mining based | active learning, data sets, empirical risk | state, model combination, show | social networks, online social, online social networks | web search, search engine, search engines |
| **ET2-tf** | commercially successful distributed data stream mining systems, previous data mining based detection systems, existing data stream management systems | empirical study, extensive empirical study, active learning | state, art model combination methods, art supervised approaches | social networks, online social networks, online social | web search, users, improve web search user browsing information |
| **ET3-tf** | data mining, detection systems, based detection systems | empirical study, extensive empirical study, active learning | state, art model combination methods, art supervised approaches | social networks, online social networks, online social | web search, users, web pages |
| **ET1-mdeg** | data mining, system, mining based | active learning, data sets, empirical risk | state, model combination, show | social networks, online social, network | web search, user, search engine |
| **ET1-medeg** | data mining, system, mining based | active learning, data sets, empirical risk | model combination, state, combination methods | social networks, online social, social network | web search, search engine, user |
| **ET1-deg/tf** | widely adopted distributed vehicle performance data mining system, minefleet distributed vehicle performance data mining system designed, community | batch mode active learning empirical risk minimization, practical batch mode active learning method, dozen benchmark real data sets verifies | existing supervised text segmentation approaches, logical shapelets significantly outperform, significantly reduce | preserving large distribute social sensor system, general news media sites acting, efficient model parameter fitting technique | controlled user study comparing generated preference judgments, helping search engine developers design micro, applications including automating repetitive search tasks |

| | | | | | |
|---|---|---|---|---|---|
| | generation module develops spatial clustering techniques | | classification error | | |
| **ET2-deg/tf** | mining based, high performance, data mining | empirical risk, active learning, active exploration | combination methods, combination approaches, supervised approaches | media sites, networking sites, online media | study user, web data, sponsored search |
| **ET3-deg/tf** | successful large, large distributed, high performance | mode active, empirical risk, active learning | segmentation approaches, art model, combination approaches | media data, massive online, news media | user browsing, enriched user, search experience |
| **EP1-tf** | data mining, distributed data mining, machine learning | parallel optimization, web search, hierarchical classification | classification, transfer learning, decision tree | social networks, twitter, diffusion of innovations | web search, clustering, implicit feedback |
| **EP2-tf** | Machine learning, experimentation, design | Web search, measurement, data mining | Decision tree, data mining, classification | twitter, social networks, data mining | clustering, experimentation, data mining |
| **EP3-tf** | mining, data, learning | learning, classification, selection | classification, learning, text | networks, social, social networks | query, search, clustering |
| **EP1-mdeg** | content systems, hierarchical summarization, file content | quantification, parallel optimization, web search | classification, decision tree, data mining | multifractal, location-based services, twitter | web search, suffix tree, search log |
| **EP2-mdeg** | security, file content, information search and retrieval | quantification, design, parallel optimization | optimization, classification, decision tree | Web search, multifractal, location-based services | information search and retrieval, web search, suffix tree |
| **EP1-deg/tf** | semi-parametric model for learning from graph, data mining for social good, vehicle data stream mining | mp k linearization, sparse structured learning, distance metric learning | zero-suppressed binary decision diagrams, generalized linear model, kernel density estimation | user generated content (ugc), user interest modeling, optimal retail location | intentional surfer model, global document analysis, heterogeneous information network |

| | | | | | |
|---|---|---|---|---|---|
| **EP2-deg/tf** | semi-parametric model for learning from graph, data mining for social good, vehicle data stream mining | information search and retrieval, classifier design and evaluation, graphs and networks | zero-suppressed binary decision diagrams, spatial databases and gis, generalized linear model | user generated content (ugc), user generated content, user interest modeling | information search and retrieval, biology and genetics, heterogeneous information network |
| **EP3-deg/tf** | winnow, computational, disaster | selection, learning, classification | cleaning, kernel, emerging | learning, computation, moments | differential, feedback, aspect |

| | | | | | |
|---|---|---|---|---|---|
| **Tópico** | framework, algorithm, information, algorithms, propose | system, deployed, automated, time, development | matching, traditional, approach, results, domain | model, models, bayesian, inference, effectiveness | based, future, location, predict, mobile |
| **Rótulos** | | | | | |
| **ET1-tf** | clustering algorithms, propose, proposed | system, monitoring systems, deployed | schema matching, correlation mining, approach | topic models, model, models kd | location, experimental results, based |
| **ET2-tf** | proposed algorithms, algorithms proposed, clustering algorithms | time monitoring systems, system, monitoring systems | correlation mining approach, computing system management traditional approaches, mining approach | structured hierarchical models kd ll, topic models inference, models inference | experimental results, location, location predictor |
| **ET3-tf** | proposed algorithms, algorithms proposed, clustering algorithms | monitoring systems, time monitoring systems, system | schema matching, correlation mining, approach | structured hierarchical models, structured hierarchical models kd ll, topic models | experimental results, location, location predictor |
| **ET1-mdeg** | propose, clustering algorithms, algorithms | system, time, monitoring systems | approach, schema matching, correlation mining | model, topic models, topic model | location, based, experimental results |
| **ET1-medeg** | clustering algorithms, propose, algorithms | system, monitoring systems, time | schema matching, approach, correlation mining | model, topic models, topic model | location, experimental results, based |
| **ET1-deg/tf** | algorithm achieves significantly higher | monitoring massive time series | tackle multiple integration | nonparametric bayes model | mobile sensors constantly probing, |

| | | | | | |
|---|---|---|---|---|---|
| | accuracy, normalized cut spectral clustering algorithms, handwriting recognition data set | moments, office development data repository, involves complex production processes | problems traditionally solve, query interfaces generally form complex matchings, computing system management traditional approaches | traditional data mining techniques, training documents requires approximate inference techniques, based inference method produces results similar | supervised learning approach based, learning spatial density models |
| **ET2-deg/tf** | data sets, data mining, knowledge information | time series, monitoring systems, product development | traditional search, holistic matching, correlation mining | kernel models, hierarchical dirichlet, regression based | based data, based social, based inference |
| **ET3-deg/tf** | algorithm achieves, log data, data dimension | production processes, development data, office development | traditionally solve, problems traditionally, mining system | model traditional, inference techniques, structured hierarchical | gps mobile, mobile sensor, mobile sensors |
| **EP1-tf** | semi-supervised clustering, bregman divergences, kernel learning | data mining, visualization, multimedia apllications | data integration, deep web, schema matching | clustering, inference, classification | trajectory patterns, spatio-temporal data mining, structural inference |
| **EP2-tf** | kernel, data mining, learning | Multimídia applications, data mining, experimentation | Schema matching, experimentation, data mining | Data integration, clustering, markov blanket | trajectory, data mining, spatial databases and gis |
| **EP3-tf** | clustering, learning, semi-supervised | data, visualization, mining | data, web, integration | clustering, inference, classifier | trajectory, data, mining |
| **EP1-mdeg** | kernel learning, dual-tree branch-and-bound, mutual reinforcement | multimedia apllications, unsupervised learning, similarity | regret minimization, text mining, web search | inference, classification, markov blanket | structural inference, smart meter data mining, positive and unlabeled learning |
| **EP2-mdeg** | kernel learning, information search and retrieval, dual-tree branch-and-bound | special architectures, multimedia apllications, | regret minimization, text mining, web search | design, inference, classifier design and evaluation | structural inference, smart meter data mining, positive and unlabeled learning |

| | | | | | |
|---|---|---|---|---|---|
| | | information search and retrieval | | | |
| **EP1-deg/tf** | hidden markov random fields, taylor series expansion, kernel k -means | bill of materials, road asset managment, part number mapping | heterogeneous translated hashing, conditional random field, cross document summarization | alerts and incidents management, block diagonal similarity matrix, clustering with dimensionality reduction | positive and unlabeled learning, smart meter data mining, spatio-temporal data mining |
| **EP2-deg/tf** | hidden markov random fields, information search and retrieval, taylor series expansion | classifier design and evaluation, information search and retrieval, bill of materials | classifier design and evaluation, information search and retrieval, on-line information services | alerts and incidents management, block diagonal similarity matrix, feature evaluation and selection | algorithm design and analysis, spatial databases and gis, positive and unlabeled learning |
| **EP3-deg/tf** | kernel, taylor, projection | event, motif, learning | event, inference, regret | neighbors, kernel, selection | and, taxi, dilemma |

| **Tópico** | mining, patterns, pattern, frequent, set | topic, text, document, documents, topics | users, recommendation, user, systems, collaborative | general, show, based, types, recently | training, labeled, classification, classifier, supervised |
|---|---|---|---|---|---|
| **Rótulos** | | | | | |
| **ET1-tf** | frequent patterns, sequential pattern, discovering frequent | topic models, latent topics, text documents | collaborative filtering, recommender systems, users | paper, data, data show | active learning, labeled data, labels |
| **ET2-tf** | frequent patterns, synthetic datasets, discover maximal frequent patterns | topic models, latent topics, text documents | collaborative filtering, recommender systems, users | data show, paper, model data | labeled data, labels, unlabeled data |
| **ET3-tf** | frequent patterns, frequent topological patterns, discovering frequent | topic models, latent topics, text documents | collaborative filtering, recommender systems, users | data show, paper, model data | labeled data, active learning, labels |
| **ET1-mdeg** | frequent patterns, sequential pattern, patterns | topic models, model, models | collaborative filtering, recommender systems, users | data, paper, based | active learning, label, labeled data |

| | | | | | |
|---|---|---|---|---|---|
| **ET1-medeg** | frequent patterns, sequential pattern, frequent topological | topic models, model, topic model | collaborative filtering, recommender systems, users | data, paper, based | active learning, labeled data, label |
| **ET1-deg/tf** | mining high utility sequential patterns sequential pattern mining plays, sequential pattern model called mutable permutation patterns, blosom efficiently mines frequent boolean expressions | information retrieval model called latent interest semantic map, hierarchical topic models disambiguating entity references, called contextual probabilistic latent semantic analysis | reputable user posts high quality comments, comment rating environment reputable users, commercial netflix cinematch recommender system | expertise difference based routing pattern, baseline topic model algorithm plsa, major asian stock market show | guided learning completely dominate smart, direct query construction active learning, show combines implicit active learning |
| **ET2-deg/tf** | regular itemsets, pattern model, data mining | latent semantic, context information, called topic | tag recommendation, recommendation based, filtering methods | based analysis, predictive models, based life | classification algorithm, binary classification, learning algorithm |
| **ET3-deg/tf** | mining plays, mining high, data mining | retrieval model, latent interest, context information | user preferences, user posts, reputable user | predictive models, evaluation showed, factor models | fully supervised, classification algorithm, label efficient |
| **EP1-tf** | sequential pattern mining, , classification | text clustering, topic models, text mining | recommender systems, collaborative filtering, tensor factorization | heterogeneous information network, clustering, text mining | active learning, cost-sensitive learning, data mining |
| **EP2-tf** | classification, data mining, experimentation | Topic models, experimentation, data mining | Tensor factorization, experimentation, data mining | Text mining, data mining, algorithms | Cost-sensitive learning, active learning, algorithms |
| **EP3-tf** | pattern, mining, data | text, topic, clustering | recommendation, filtering, recommender | clustering, data, learning | learning, classification, active |
| **EP1-mdeg** | sequential pattern mining, , classification | context, text mining, lda | ensemble learning, recommendation, | text mining, statistical topic models, heterogeneou | cost intervals, classification, cost-sensitive learning |

| | | | | | |
|---|---|---|---|---|---|
| | | | hierarchical smoothing | s information network | |
| **EP2-mdeg** | sequential pattern mining, , classification | human factors, context, text mining | ensemble learning, information search and retrieval, recommendation | design, text mining, statistical topic models | information search and retrieval, cost intervals, classifier design and evaluation |
| **EP1-deg/tf** | high utility sequential pattern mining, closed and free itemsets, sequential pattern mining | latent semantic indexing, latent dirichlet allocation, partially supervised learning | hybrid content and collaborative filtering, method of moments estimation, popularity based performance evaluatoin | coupled hidden markov model, hidden markov models, coupled behavior analysis | interactive and online data mining, support vector machines, sample selection bias |
| **EP2-deg/tf** | graph and tree search strategies, high utility sequential pattern mining, closed and free itemsets | information search and retrieval, latent semantic indexing, latent dirichlet allocation | user profiles and alert services, hybrid content and collaborative filtering, method of moments estimation | information search and retrieval, spatial databases and gis, coupled hidden markov model | interactive and online data mining, information search and retrieval, classifier design and evaluation |
| **EP3-deg/tf** | and, stream, concise | em, lda, filtering | naïve, and, filtering | statistical, classification, computing | and, misclassification , selection |

| | | | | | |
|---|---|---|---|---|---|
| **Tópico** | problem, algorithm, optimal, algorithms, solution | detection, streams, time, detect, real | community, communities, interactions, properties, demonstrate | entities, knowledge, example, extracted, information | models, model, maximum, markov, structure |
| **Rótulos** | | | | | |
| **ET1-tf** | approximation algorithm, approximation algorithms, optimization problem | data streams, data stream, spam detection | community structure, communities, identifying communities | natural language, knowledge bases, language text | hidden markov models, hidden markov, markov models |

| | | | | | |
|---|---|---|---|---|---|
| **ET2-tf** | approximation algorithms, approximation algorithm, algorithms | data streams, realize topic emergence detection, change detection | community structure, identifying communities, communities | natural language, natural language text, language text | hidden markov models, markov models, models |
| **ET3-tf** | approximation algorithms, approximation algorithm, optimization problem | data streams, spam detection, review spam detection | community structure, identifying communities, communities | natural language, natural language text, language text | hidden markov models, markov models, hidden markov |
| **ET1-mdeg** | algorithm, algorithms, problem | data streams, data stream, time | community structure, communities, community | natural language, knowledge, information | hidden markov, model, markov models |
| **ET1-medeg** | approximation algorithm, algorithm, algorithms | data streams, data stream, spam detection | community structure, communities, community | natural language, knowledge, knowledge bases | hidden markov, markov models, hidden markov models |
| **ET1-deg/tf** | optimization problem called {\em seed minimization, show interesting general theoretical properties, nuclear norm minimization problem | data stream mining faces hard constraints, proposed algorithm produces classification accuracy comparable, drifting data streams data stream mining | community profiling model called cocomp, world large scale networks demonstrate, previously studied network properties | probabilistic knowledge fusion recent years, previously published structured knowledge repository, scale probabilistic knowledge base | guided conditional random field learning conditional random fields, markov chain monte carlo procedure based, world data sets involving email communication |
| **ET2-deg/tf** | minimization problems, general problem, approximation guarantee | stream learning, stream classification, data stream | structure accurately, changing community, explanatory community | text processing, repository information, scale knowledge | conditional random, random fields, hidden markov |
| **ET3-deg/tf** | minimization problem, minimization problems, problem called | stream mining, stream classification, realize topic | called cocomp, model called, community profiling | structured knowledge, knowledge fusion, probabilistic knowledge | random field, conditional random, maximum discriminative |

| | | | | | |
|---|---|---|---|---|---|
| **EP1-tf** | feature selection, approximation algorithms, proportional fault tolerance | data streams, concept drift, closed mining | evolution, interaction networks, multi-mode networks | text mining, machine learning, version history | classification, gibbs sampling, hidden markov model |
| **EP2-tf** | Approximation algorithms, data mining, miscellaneous | Closed mining, data mining, data streams | Multi-mode networks, data mining, measurement | Version history, data mining, experimentation | Gibbs sampling, data mining, experimentation |
| **EP3-tf** | recommendation, learning, selection | data, concept, learning | networks, evolution, dynamic | text, clustering, recommendation | classification, hidden, markov |
| **EP1-mdeg** | approximation algorithms, proportional fault tolerance, item design | ensemble learning, spatial indexing, ensemble methods | blog, multi-mode networks, neighborhood analysis | text mining, version history, developer expertise | em, predictive, classification |
| **EP2-mdeg** | miscellaneous, approximation algorithms, proportional fault tolerance | ensemble learning, spatial indexing, classifier design and evaluation | blog, multi-mode networks, neighborhood analysis | text mining, version history, developer expertise | em, predictive, information search and retrieval |
| **EP1-deg/tf** | adaptive line search scheme, regularized least squares classification, dynamic social networks | graph signal processing, graph signal filtering, support vector machines | transpose closures from a bipartite graph database, diffusion of innovations, contact group clustering | probabilistic generative model, connecting the dots, job matching system | conditional random fields, extended saddle points, hidden markov model |
| **EP2-deg/tf** | parallel and vector implementations, adaptive line search scheme, regularized least squares classification | classifier design and evaluation, graph signal processing, graph signal filtering | transpose closures from a bipartite graph database, group and organization interfaces, diffusion of innovations | information search and retrieval, probability and statistics, connecting the dots | information search and retrieval, biology and genetics, probability and statistics |
| **EP3-deg/tf** | nesterov's, kernel, anti | concept, methods, stream | evolutionary, neighborhood, classification | redescriptions, storytelling, intelligence | em, upper, filters |

| | | | | | |
|---|---|---|---|---|---|
| **Tópico** | high, results, feature, experimental, features | graph, graphs, network, nodes, networks | analysis, dimensional, consider, space, subset | privacy, preserving, information, data, private | time, series, events, temporal, event |
| **Rótulos** | | | | | |
| **ET1-tf** | experimental results, visual features, results | social networks, social network, nodes | matrix, finding, subset | data mining, preserving data, data collection | time series, time series knowledge, series knowledge |
| **ET2-tf** | experimental results, visual features, features | social networks, large social networks graphs, nodes | dimensional space, dimensional spaces, matrix | data mining, preserving data, preserving data mining | time series knowledge mining temporal patterns composed, time series, news events |
| **ET3-tf** | experimental results, visual features, features | social networks, large social networks graphs, social networks graphs | dimensional space, dimensional spaces, matrix | preserving data, data mining, preserving data mining | time series, time series knowledge mining temporal patterns composed, time series knowledge |
| **ET1-mdeg** | experimental results, results, visual features | social networks, social network, graph | matrix, finding, space | data mining, data, preserving data | time series, time, series knowledge |
| **ET1-medeg** | experimental results, visual features, results | social networks, social network, graph | matrix, finding, high dimensional | data mining, preserving data, data | time series, time, series knowledge |
| **ET1-deg/tf** | deliver semantically relevant image retrieval, art tensor factorization methods, simultaneous tensor subspace selection | real datasets demonstrates superior performance, social communication network social networks, pure social network topology | high dimensional robust correlation, significantly reduced computational cost, interpretable nonnegative matrix decompositions | multiple organizations independently release anonymized data, location privacy protection methods based, resistant anonymous data collection method | production equipment monitoring system learning temporal graph structures, based data mining approaches overlook dynamic features, time series data reveals important dependency relationships |
| **ET2-deg/tf** | subspace selection, image feature, model selection | large network, social network, large networks | high dimensional, high computational, computational cost | preserving distributed, distributed data, preserving protocol | patterns based, temporal data, learning temporal |

| | | | | | |
|---|---|---|---|---|---|
| **ET3-deg/tf** | gene selection, semantically relevant, relevant image | social communication, communication network, compressing social | dimensional robust, reduced computational, high dimensional | preserving protocol, data release, preserving distributed | patterns composed, patterns based, mining temporal |
| **EP1-tf** | singular value decomposition, noise removal, clusranking | social networks, approximate algorithm, community detection | consecutive ones property, random projections, pincipal component analysis | privacy, data mining, anonymity | interval patterns, surveillance systems, structure learning |
| **EP2-tf** | Noise removal, data mining, image databases | Social networks, data mining, theory | Random projections, data mining, experimentation | security, interval pattern, data mining | Structure learning, data mining, algorithms |
| **EP3-tf** | data, image, mining | graph, clustering, proximity | data, random, sampling | privacy, anonymity, anonymization | graph, dynamic, time |
| **EP1-mdeg** | singular value decomposition, noise removal, clusranking | proximity, escape probability, approximate algorithm | clustering, pincipal component analysis, banded matrices | proximity, classification, data publishing | interval patterns, surveillance systems, structure learning |
| **EP2-mdeg** | singular value decomposition, information search and retrieval, database applications | proximity, escape probability, graph algorithms | pincipal component analysis, banded matrices, data mining | clustering, classification, data publishing | interval patterns, learning, surveillance systems |
| **EP1-deg/tf** | singular value decomposition, principal component analysis, meaningful itemset mining | strength of weak ties, kernel on a graph, learning to rank | probabilistic latent semantic analysis, pincipal component analysis, consecutive ones property | theta-secure cloaking area, secure multiparty computation, data publishing | incremental singular value decomposition, exponential random graph model, graph rewriting rule |
| **EP2-deg/tf** | information search and retrieval, | graph and tree search strategies, | probabilistic latent semantic analysis, | spatial databases and gis, theta-secure cloaking | incremental singular value decomposition, information search and |

| | | | | | |
|---|---|---|---|---|---|
| | feature evaluation and selection, singular value decomposition | strength of weak ties, kernel on a graph | nonnumerical algorithms and problems, pincipal component analysis | area, secure multiparty computation | retrieval, content analysis and indexing |
| **EP3-deg/tf** | em, kernel, selection | and, kernel, filtering | nearest, presence/absence, semantic | differential, secure, classification | invariance, warping, differential |

| **Tópico** | paper, world, real, tasks, mining | clustering, cluster, clusters, objects, experiments | online, advertising, real, scale, ad | time, large, faster, experiments, fast | product, describe, products, decision, online |
|---|---|---|---|---|---|
| **Rótulos** | | | | | |
| **ET1-tf** | data mining, paper, data sets | clusters, clustering, clustering algorithms | display advertising, guaranteed display advertising, guaranteed display | large graphs, experiments, 8 times faster | products, customers, customer reviews |
| **ET2-tf** | world data mining applications, data mining, world applications | clusters, clustering, clustering algorithms | display advertising, guaranteed display advertising, online display advertising | 8 times faster, times faster, large graphs efficiently | products, customers, summarizing customer reviews merchants selling products |
| **ET3-tf** | data mining, world data mining applications, world data | clusters, clustering, clustering algorithms | display advertising, guaranteed display advertising, online advertising | 8 times faster, times faster, large graphs efficiently | products, customers, selling products |
| **ET1-mdeg** | data mining, data, data sets | clustering, clusters, cluster | display advertising, ad, display ad | fast, large, time | products, product, customer reviews |
| **ET1-medeg** | data mining, data, data sets | clustering, clusters, cluster | display advertising, display ad, ad | fast, large, time | products, product, customer reviews |
| **ET1-deg/tf** | deployed data mining application system, deployed data mining system, naturally handle | approach seamlessly integrates heterogenous data types measured, categorical data clustering categorical data | greedy mechanism high performing advertisers tend, data mining applications including computational advertising, online ad | fast parallel graph engine handling billion, training linear support vector machines, limited memory | ford motor company obtained practical experience, puzzling outcomes explained online controlled experiments, summarizing customer |

| | | | | | |
|---|---|---|---|---|---|
| | multiple outputs | poses, real world data sets validate | exchange marketplace display advertising | maximal clique enumeration | reviews merchants selling products |
| **ET2-deg/tf** | world power, data sets, data mining | real data, categorical objects, real world | ad exchange, ad allocation, world system | limited memory, linear svm, large real | customer reviews, prediction market, online controlled |
| **ET3-deg/tf** | mining system, mining application, deployed data | real data, quality hierarchical, approach seamlessly | ad exchange, marketplace display, scale data | fast parallel, training linear, linear support | summarizing customer, practical experience, customer reviews |
| **EP1-tf** | classification, k-means distance, cost-sensitive learning | clustering, minimum description length, minimum description length principle | display advertising, internet advertising, budget allocation | massive networks, distributed computing, i/o efficient | sentiment analysis, text mining, opinion mining |
| **EP2-tf** | Cost-sensitive learning, data mining, algorithms | Minimum description length, data mining, clustering | Budget allocation, experimentation, display advertising | computing, data mining, experimentatio n | analysis, data mining, experimentation |
| **EP3-tf** | learning, mining, clustering | clustering, data, description | advertising, display, display advertising | graph, distributed, graphs | mining, sentiment, classification |
| **EP1-mdeg** | supervised projection, classification, k-means distance | clustering, episode mining, coding costs | internet advertising, budget allocation, hierarchy | i/o efficient, sparse graphs, sampling | forecasting, classification, mobile recommender systems |
| **EP2-mdeg** | supervised projection, linked representations , classification | clustering, episode mining, coding costs | theory, internet advertising, budget allocation | information search and retrieval, i/o efficient, sparse graphs | efficient, forecasting, classification |
| **EP1-deg/tf** | principal component analysis, class association rules, heterogeneous social network | minimum description length principle, minimum description length, normalized maximum likelihood | spike and slab prior, spars contingency tables, bid landscape forecasting | latent dirichlet allocation, succinct data structure, locality sensitive hashing | mobile recommender systems, market share rules, statistical quantitative rules |
| **EP2-deg/tf** | feature evaluation and selection, | minimum description length principle, minimum | computations on discrete structures, spike and slab prior, | information search and retrieval, | mobile recommender systems, market share |

| | | | | | |
|---|---|---|---|---|---|
| | principal component analysis, time series analysis | description length, normalized maximum likelihood | language parsing and understanding | classifier design and evaluation, latent dirichlet allocation | rules, statistical quantitative rules |
| **EP3-deg/tf** | selection, projection, classification | kernel, selection, series | and, marketplace, computational | coefficient, computing, search | insurance, hmms, forecasting |

| **Tópico** | work, design, large, scale, evaluate | approach, single, simple, key, test |
|---|---|---|
| **Rótulos** | | |
| **ET1-tf** | large scale, large scale visual, scale visual | paper, hypothesis testing, statistical testing outlier |
| **ET2-tf** | large scale visual recommendations, algorithms, completely automated large scale visual recommendation system | statistical testing outlier detection, hypothesis testing, paper |
| **ET3-tf** | large scale, large scale visual, scale visual | statistical testing outlier detection, statistical testing, statistical testing outlier |
| **ET1-mdeg** | work, scale visual, scale | test, paper, testing |
| **ET1-medeg** | scale visual, large scale, work | test, paper, testing outlier |
| **ET1-deg/tf** | completely automated large scale visual recommendation system, existing cardinal peer grading methods, large scale visual recommendations | develop simple sufficient conditions, natural language processing techniques, supports large permutation test |
| **ET2-deg/tf** | scale data, worker quality, existing methods | statistical model, statistical inferences, make statistical |
| **ET3-deg/tf** | automated large, scale visual, existing cardinal | testing outlier, natural language, develop simple |
| **EP1-tf** | social networks, revenue optimisation, record linkage | classification, networks, |
| **EP2-tf** | networks, data mining, information search and retrieval | networks, data mining, performance |
| **EP3-tf** | networks, graph, personalization | networks, classification, learning |
| **EP1-mdeg** | revenue optimisation, record linkage, taxonomy building | networks, filtering, classification |
| **EP2-mdeg** | networks, information search and retrieval, revenue optimisation | networks, filtering, classification |
| **EP1-deg/tf** | latent dirichlet allocation, bayesian rose tree, opinion aspect extraction | automated generation of visualizatons, approximate subgraph isomorphism, association rule mining |
| **EP2-deg/tf** | content analysis and indexing, information search and retrieval, security and protection | automated generation of visualizatons, biology and genetics, approximate subgraph isomorphism |

| | | |
|---|---|---|
| **EP3-deg/tf** | control, inference, aspect | filtering, transductive, learning |

# APÊNDICE D – RESULTADOS SDC

| **Tópico** | signal, noise, filter, signals, estimation | students, university, course, education, science | method, proposed, results, algorithm, adaptive | network, networks, services, mobile, service | query, queries, database, databases, relational |
|---|---|---|---|---|---|
| **Rótulos** | | | | | |
| **ET1-tf** | noise signals, signal processing, fourier transform | computer science, power engineering, power engineering courses | experimental results, game algorithm, algorithm | radio access, networked multimedia, umts terrestrial radio | xml documents, processing xml, processing xml queries |
| **ET2-tf** | noise signals, valued signals, cryptographic applications exploiting nonlinear signal processing | computer science courses, science courses, computer science | experimental results, experimental results show, results show | rich networked multimedia future, umts terrestrial radio access network, based umts terrestrial radio access network | processing xml queries, relational databases, processing xml |
| **ET3-tf** | noise signals, valued signals, exploiting nonlinear signal | power engineering courses, engineering courses, computer science courses | experimental results, experimental results show, results show | networked multimedia, access network, radio access network | processing xml queries, relational databases, processing xml |
| **ET1-mdeg** | signal processing, fourier transform, signal | computer science, power engineering, engineering courses | algorithm, experimental results, game algorithm | radio access, terrestrial radio, umts terrestrial | xml documents, xml queries, processing xml |
| **ET1-medeg** | signal processing, fourier transform, applications exploiting | computer science, power engineering, engineering courses | algorithm, experimental results, game algorithm | radio access, terrestrial radio, umts terrestrial | xml documents, xml queries, processing xml |
| **ET1-deg/tf** | cryptographic applications exploiting nonlinear signal processing, | professional development project involving louisiana state university, incorporating computer based | fuzzy association rules mining algorithm, power flow tracking methods based, face | based umts terrestrial radio access network, umts terrestrial radio access network, qos | signal processing fast hadamard transform, motivate clustering relational proximity data, |

| | | | | | |
|---|---|---|---|---|---|
| | colored noise rotated rectangular symbol constellations, image processing applications digital signal | multimedia material, enhance undergraduate/graduate power engineering courses | detection algorithm based | ip+atm switch router architecture | signal processing applications |
| **ET2-deg/tf** | fourier transform, orthogonal frequency, noise analysis | power engineering, enhance teaching, learning projects | based algorithm, improved linearization, improved maisheng | access optimization, networked multimedia, ip qos | signal processing, proximity relational, relational data |
| **ET3-deg/tf** | exploiting nonlinear, signal processing, nonlinear signal | state university, enhance teaching, power engineering | tracking methods, mining algorithm, methods based | radio access, access network, ip+atm switch | processing fast, signal processing, relational data |
| | | | | | |
| **Tópico** | verification, model, formal, specification, checking | optimal, problem, optimization, algorithm, minimize | memory, cache, reduce, access, overhead | networks, wireless, network, nodes, ad | book, concepts, introduction, examples, computer |
| **Rótulos** | | | | | |
| **ET1-tf** | modeling mpi, declarative models, existing neurone models | optimization problem, problem, layer optimization | trace cache, block cache, based trace cache | initiates thread, independent reading, taking advantage | book, computer, books |
| **ET2-tf** | modeling mpi, declarative models, existing neurone models | optimization problem, design approximation algorithms, approximation algorithms | trace cache, block cache, encapsulation scheme | initiates thread, independent reading, taking advantage | book, computer, books |
| **ET3-tf** | modeling mpi, declarative models, | optimization problem, design approximation algorithms, | trace cache, block cache, based trace cache | initiates thread, independent | book, computer, books |

| | | | | | |
|---|---|---|---|---|---|
| | existing neurone models | approximation algorithms | | reading, taking advantage | |
| **ET1-mdeg** | model, modeling, neurone models | problem, optimization problem, location problem | trace cache, based trace, block cache | read, initiates thread, broadcasting systems | book, computer, books |
| **ET1-medeg** | model, neurone models, existing neurone | problem, optimization problem, location problem | trace cache, based trace, block cache | read, initiates thread, broadcasting systems | book, computer, books |
| **ET1-deg/tf** | smale real number model, shared memory model based, global model creation | highly efficient multiobjective evolutionary algorithm, cloud computing resource allocation problem, multiobjective euclidean location problem | based trace cache renames fetch addresses, based trace cache implementation, level fault diagnosis scheme | time broadcasting systems, initiates thread relocation, traditionally relied heavily | locate related topics quickly, edge approach researchers, academic topics |
| **ET2-deg/tf** | memory systems, global model, model creation | efficient solutions, resource constraints, search algorithm | trace cache, fetch schemes, block cache | traditional expert, traditionally relied, gradually combining | approach researchers, academic topics, personal computers |
| **ET3-deg/tf** | number model, model based, memory model | efficient multiobjective, multiobjective evolutionary, multiobjective euclidean | cache renames, fetch schemes, cache achieves | traditional expert, traditionally relied, gradually combining | related topics, topics quickly, approach researchers |
| | | | | | |
| **Tópico** | position, vehicle, navigation, robot, mobile | network, networks, traffic, packet, delay | bound, bounds, complexity, lower, upper | models, performance, evaluate, accurate, model | al, et, 10, 12, 11 |
| **Rótulos** | | | | | |
| **ET1-tf** | superposition strategy, | wireless networks, ofdma networks, area networks | problem, upper bound, local polynomial | make inaccurate absolute, make inaccurate, | local search, proposed |

| | | | | | |
|---|---|---|---|---|---|
| | mobile wimax, robotic vision | | | absolute performance | algorithm, cylindrical annuli |
| **ET2-tf** | superposition strategy, mobile wimax, robotic vision | wireless networks, ofdma networks, networks femtocells | upper bound, problem, local polynomial | make inaccurate absolute performance predictions, benchmark performance prediction standard benchmarking, accurately predict architectural trends | local search, theoretical perspective, 2006 special issue |
| **ET3-tf** | superposition strategy, mobile wimax, robotic vision | wireless networks, area networks, ofdma networks | upper bound, problem, local polynomial | make inaccurate absolute performance predictions, performance predictions, absolute performance predictions | local search, theoretical perspective, 2006 special issue |
| **ET1-mdeg** | superposition strategy, mobile wimax, position | network, area networks, wireless networks | problem, upper bound, polynomial | performance, performance prediction, absolute performance | local search, proposed algorithm, algorithm |
| **ET1-medeg** | superposition strategy, mobile wimax, modification inrobotic | area networks, wireless networks, network | problem, upper bound, medical computational | performance prediction, absolute performance, inaccurate absolute | local search, proposed algorithm, finite difference |
| **ET1-deg/tf** | arm trajectory modification inrobotic manipulators, multiple coupled vehicle | multihop wireless network remains largely unaddressed, 3g networks 3g cellular data networks, large scale | problem/case based learning constitutes, constructing medical computational problems, semantically annotates problems | benchmark performance prediction standard benchmarking, make inaccurate absolute | alphaserver es40 system showing high accuracy, singularly perturbed general differential difference |

| | | | | | |
|---|---|---|---|---|---|
| | systems, universal mobile telecommunications system | multicast packet switch | | performance predictions, simulators predict trends accurately | equations, jikes research java virtual machine |
| **ET2-deg/tf** | tracking system, robotic vision, time tracking | wireless networks, networks femtocells, access networks | annotates problems, biological sequence, separate problems | predict architectural, important performance, performance effects | natural convection, deterministic transport, carlo methods |
| **ET3-deg/tf** | inrobotic manipulators, modification inrobotic, coupled vehicle | data networks, networks 3g, 3g networks | problem/case based, computational problems, annotates problems | make inaccurate, inaccurate absolute, benchmark performance | alphaserver es40, differential difference, virtual machine |
| | | | | | |
| **Tópico** | 3d, rendering, graphics, interactive, visualization | security, attacks, attack, detection, network | business, technology, government, management, organizations | question, problem, answer, questions, fact | spatial, temporal, analysis, information, visualization |
| **Rótulos** | | | | | |
| **ET1-tf** | image based rendering, virtual environments, image based | likelihood sequence, internet, detection algorithms | laboratory information, sector information systemissues, public sector | boolean satisfiability, polynomial cases, satisfiability problems | analysis, techniques adapted, improved spatial |
| **ET2-tf** | image based rendering, virtual environments, image based | detection algorithms, likelihood sequence detection, fraud attacks | information management, information services, laboratory information services | polynomial cases, simple polynomial cases, combinatorial complexity problems | analysis, techniques adapted, improved spatial |
| **ET3-tf** | image based rendering, virtual environments, image based | detection algorithms, likelihood sequence detection, fraud attacks | information management, information services, laboratory information services | satisfiability problems, boolean satisfiability problems, polynomial cases | analysis, techniques adapted, improved spatial |

| | | | | | |
|---|---|---|---|---|---|
| **ET1-mdeg** | image based, based rendering, virtual environments | likelihood sequence, internet, detect | information, laboratory information, information system | boolean satisfiability, satisfiability problems, problem | analysis, improved spatial, spatial distribution |
| **ET1-medeg** | image based, based rendering, virtual environments | likelihood sequence, internet, detect | laboratory information, information system, information | boolean satisfiability, satisfiability problems, boolean satisfiability problems | analysis, improved spatial, spatial distribution |
| **ET1-deg/tf** | real world imagery demonstratet he validity, residual error image, image based rendering | defeasible security policy composition, proactively detect automated traffic, proposed detection algorithms | leading business school doesnâ__t, york times business section, interorganisational public policy implementation | partite boolean satisfiability problems, bipartite boolean satisfiability problem, boolean satisfiability problems | improved spatial distribution, techniques adapted, information extraction |
| **ET2-deg/tf** | based rendering, image based, error image | sequence detector, security breach, based security | public computer, public infrastructures, laboratory information | satisfiability problems, complexity problems, polynomial cases | improved spatial, spatial distribution, techniques adapted |
| **ET3-deg/tf** | imagery demonstratet he, world imagery, based rendering | defeasible security, security policy, proactively detect | business school, leading business, public policy | satisfiability problems, satisfiability problem, complexity problems | improved spatial, spatial distribution, techniques adapted |
| | | | | | |
| **Tópico** | coding, compression, video, image, scheme | wiley, periodicals, john, copyright, sons | learning, knowledge, actions, agent, intelligence | user, interface, interaction, users, interfaces | computing, resources, distributed, service, cloud |
| **Rótulos** | | | | | |
| **ET1-tf** | image, quality improvement, schemes | research, research work, research shows | artificial intelligence, real estate, estate transactions | users, human visual, visual system | resources, cloud computing technologies, multimedia services |

| | | | | | |
|---|---|---|---|---|---|
| **ET2-tf** | image, separated, quality improvement | research, research work, research shows | estate transactions, artificial intelligence, real estate transactions | users, human visual, visual system | resources, multimedia services, applying cloud computing technologies |
| **ET3-tf** | image, separated, quality improvement | research, research work, research shows | artificial intelligence, estate transactions, real estate transactions | users, human visual, visual system | resources, cloud computing technologies, multimedia services |
| **ET1-mdeg** | image, quality, generate | research, research work, research shows | real estate, estate transactions, artificial intelligence | users, human visual, visual system | resources, cloud computing, applying cloud |
| **ET1-medeg** | image, viewpoint image, standard definition | research, research work, research shows | real estate, estate transactions, real estate transactions | users, human visual, visual system | cloud computing, resources, applying cloud |
| **ET1-deg/tf** | standard definition video, viewpoint image acquisition, video data | research work, research shows, research | robust interaction control algorithm, spoken language interaction, massively multiagent system | human visual system, social interactions, optimize user | generalized gamma distributed call holding times, creative computing research area, applying cloud computing technologies |
| **ET2-deg/tf** | definition video, video data, viewpoint images | research work, research shows, research | language interaction, estate transactions, satisfaction problems | human visual, visual system, social interactions | performance computing, subgrid modelling, scarce resource |
| **ET3-deg/tf** | definition video, video data, viewpoint images | research work, research shows, research | interaction control, robust interaction, language interaction | human visual, visual system, social interactions | gamma distributed, distributed call, computing technologies |
| | | | | | |
| **Tópico** | programming, code, web, developers, book | network, networks, nodes, peer, node | mining, patterns, discovery, algorithm, pattern | study, factors, findings, research, influence | distributed, message, communication, messages, ii |

| Rótulos | | | | | |
|---|---|---|---|---|---|
| **ET1-tf** | book, codebook search /, search / | normal neural, chaotic neural, normal neural networks | approximation algorithms, cut algorithms, heuristics optimization algorithms | intrinsic motivation, study, hygiene factors | forward neural networks minimizing, neural network algorithms, asynchronous stochastic dynamics |
| **ET2-tf** | codebook search /, codebook search, book | normal neural networks, neural networks, proactive/reactive communication | approximation algorithms, cut algorithms, heuristics optimization algorithms | study, hygiene factors, influence | forward neural networks minimizing, asynchronous stochastic dynamics, local minimizer |
| **ET3-tf** | codebook search /, codebook search, book | normal neural networks, neural networks, chaotic neural network | approximation algorithms, cut algorithms, heuristics optimization algorithms | study, hygiene factors, influences workplace attitudes | forward neural networks minimizing, neural network algorithms, asynchronous stochastic dynamics |
| **ET1-mdeg** | book, search /, codebook search | neural network, network, chaotic neural | heuristics optimization, optimization algorithms, real | intrinsic motivation, study, influence | neural network, neural networks, forward neural |
| **ET1-medeg** | book, search /, codebook search | neural network, chaotic neural, chaotic neural network | heuristics optimization, optimization algorithms, heuristics optimization algorithms | intrinsic motivation, study, influence | neural network, neural networks, forward neural |
| **ET1-deg/tf** | numerous freely downloadable codes, code excited linear prediction, | chaotic neural network constructed, normal neural networks, chaotic neural network | algorithm runs online, reality learning environment, heuristics optimization algorithms | intrinsic motivation positively influences workplace attitudes, study examines intrinsic | forward neural networks minimizing, layer neural network algorithms, asynchronous |

| | | | | | |
|---|---|---|---|---|---|
| | codebook search / | | | motivation, hygiene factors | stochastic dynamics |
| **ET2-deg/tf** | covers applications, modular code, code generation | neural networks, neural network, communication approach | reality learning, algorithm runs, optimization algorithms | hygiene factors, findings suggest, empirical examination | asynchronous stochastic, local overfitting, local minimizer |
| **ET3-deg/tf** | downloadable codes, code excited, covers applications | network constructed, neural networks, neural network | reality learning, algorithm runs, optimization algorithms | positively influences, influences workplace, study examines | neural networks, networks minimizing, neural network |
| | | | | | |
| **Tópico** | world, people, today, book, get | circuits, circuit, power, delay, chip | user, users, web, tools, interface | web, users, user, social, online | book, guide, publisher, questions, help |
| **Rótulos** | | | | | |
| **ET1-tf** | book, book explores, fully illustrated book | reversible circuits, circuits applications, prototype circuits | user interface, user interface concept, information | online learning, utilizing users, device users | oracle vm, vm manager, oracle vm manager |
| **ET2-tf** | book, book explores, fully illustrated book makes | reversible circuits, circuits applications, prototype circuits | user interface, users, innovative user interface concept | utilizing users, device users, users | book covers, book, business |
| **ET3-tf** | book, book explores, fully illustrated book | reversible circuits, circuits applications, prototype circuits | user interface, user interface concept, users | utilizing users, device users, users | book covers, book, business |
| **ET1-mdeg** | book, book explores, illustrated book | reversible circuits, circuits applications, prototype circuits | user interface, information, web | users, online learning, user | oracle vm, vm manager, learn |
| **ET1-medeg** | book, book explores, illustrated book | reversible circuits, circuits applications, reversible circuits applications | user interface, information, innovative user | online learning, users, disappoints users | oracle vm, vm manager, oracle vm manager |
| **ET1-deg/tf** | fully illustrated book makes, book full color images, book | eventual technology adopted, reversible circuits applications, reversible circuits | tecate dynamically crafts user, relevant category specific information, | class online learning algorithm, existing researches | oracle vm managersecure oracle vm managerlearn xen |

| | | | | | |
|---|---|---|---|---|---|
| | introduces teachers | | innovative user interface concept | treat, online learning problem | utilities, oracle vm manager helps administrators manage, vmwarelearn powerful xen hypervisor utilities |
| **ET2-deg/tf** | book introduces, business letter, world examples | eventual technology, technology adopted, circuits applications | user interface, user selected, mysql user | online learning, feature information, information infrastructure | vm serverslearn, study guide, covers payload |
| **ET3-deg/tf** | illustrated book, book full, book makes | eventual technology, technology adopted, circuits applications | innovative user, interface concept, crafts user | class online, online learning, feature information | vm managerlearn, managerlearn xen, learning oracle |
| | | | | | |
| **Tópico** | problem, search, problems, solving, algorithm | agent, agents, multi, distributed, systems | discussed, described, presented, given, general | optimization, genetic, algorithm, search, evolutionary | logic, semantics, reasoning, theory, notion |
| **Rótulos** | | | | | |
| **ET1-tf** | algorithm, parallel algorithms, monte carlo | agent systems, agents, autonomous agent | general principles apply, general nature, presented | evolutionary algorithms, proposed, open problems | ambient calculus, propositional proofs, classical proofs |
| **ET2-tf** | parallel algorithms, art solutions, solutions | fully autonomous agent systems, agents, command agents | general principles apply, general nature, presented | evolutionary algorithms, open problems, problems | proof semantics, ambient calculus, calculus |
| **ET3-tf** | parallel algorithms, art solutions, solutions | agent systems, autonomous agent systems, fully autonomous agent systems | general principles apply, general nature, presented | evolutionary algorithms, open problems, problems | proof semantics, ambient calculus, propositional proofs |
| **ET1-mdeg** | algorithm, algorithms, monte carlo | agent systems, agents, autonomous agent | principles apply, general principles, presented | evolutionary algorithms, algorithms, proposed | propositional proofs, ambient calculus, classical proofs |

| | | | | | |
|---|---|---|---|---|---|
| **ET1-medeg** | algorithm, monte carlo, chain monte | agent systems, autonomous agent, agents | principles apply, general principles, general principles apply | evolutionary algorithms, algorithms, open problems | propositional proofs, ambient calculus, classical proofs |
| **ET1-deg/tf** | operations research 2006 meritorious service award recipients, markov chain monte carlo algorithms, algorithm demands large computational resources | manually curated highly reliable multiple sequence alignments, learning agent employing reinforcement learning, agent technology require humans | general principles apply, general nature, communications infrastructure | make researchers aware, query distribution problem, linear time algorithm | cartesian closed categories capture intuitionistic propositional proofs, understand classical propositional proofs based, categories characterizes classical proofs |
| **ET2-deg/tf** | searching probabilistic, robust multiresolution, multiresolution hypothesis | learning agent, agent systems, multiple users | general principles, general nature, communications infrastructure | researchers aware, make researchers, problem posed | calculus \cite{par92}, propositional proofs, proofs based |
| **ET3-deg/tf** | carlo algorithms, algorithm demands, operations research | reliable multiple, multiple sequence, fully autonomous | general principles, general nature, communications infrastructure | researchers aware, make researchers, problem posed | calculus \cite{par92}, propositional proofs, intuitionistic logic |
| | | | | | |
| **Tópico** | optimization, matrix, linear, problem, algorithm | semantic, knowledge, ontology, domain, ontologies | water, environmental, energy, china, study | brain, visual, subjects, human, activity | image, images, segmentation, method, reconstruction |
| **Rótulos** | | | | | |
| **ET1-tf** | linear complementarity problems, complementarity problems, linear | domain experts, domain, semantical framework | lowlying area, china, study | greater activity, visualizes, areas showed greater | proposed, hybrid md, texture image |

| | | | | | |
|---|---|---|---|---|---|
| | complementarity | | | | |
| **ET2-tf** | linear complementarity problems, complementarity problems, methods solve linear partial differential equations | domain experts, domain, domain adaptive model based | urbanized lowlying area, areas, lowlying area | visual working memory studies, visualization, localizing sounds recruited greater activity | texture image segmentation, texture images demonstrate, images |
| **ET3-tf** | linear complementarity problems, complementarity problems, methods solve linear | domain experts, domain, semantical framework | urbanized lowlying area, areas, lowlying area | greater activity, visualizes, nonspatial auditory tasks preferentially recruit dorsal | texture image segmentation, texture images demonstrate, images |
| **ET1-mdeg** | complementarity problems, linear complementarity, problem | web, domain, meaningful metadata | lowlying area, china, study | greater activity, activity, auditory task | method, proposed, hybrid md |
| **ET1-medeg** | complementarity problems, linear complementarity, linear complementarity problems | meaningful metadata, adaptive model, web meaningful | lowlying area, china, economic losses | greater activity, auditory task, activity | hybrid md, texture image, method |
| **ET1-deg/tf** | methods solve linear partial differential equations, multiprocessor cache coherence problem vanishes, interior point methods | wide web meaningful metadata describing, domain adaptive model based, metadata description standards | liaoning coastal highway area, increased economic losses compared, urbanized lowlying area | lateral superior parietal areas showed greater activity, nonspatial auditory tasks preferentially recruit dorsal, visual environment explorer supports developers | pareto local search method applied, bestselling author philip andrews reconstruction, hybrid md coding method |

| | | | | | |
|---|---|---|---|---|---|
| **ET2-deg/tf** | proposes methods, point methods, evaluate methods | metadata description, description standards, common semantical | economic losses, lowlying area, urbanized lowlying | brain areas, hemodynamic activity, auditory task | proposed protocol, texture segmentation, texture image |
| **ET3-deg/tf** | linear partial, methods solve, solve linear | wide web, web meaningful, metadata description | increased economic, highway area, economic losses | greater activity, tasks preferentially, visual environment | method applied, md method, search method |
| | | | | | |
| **Tópico** | oriented, model, models, modeling, development | estimation, probability, distribution, model, models | social, theory, human, research, science | optical, measurement, surface, light, laser | research, recent, challenges, attention, future |
| **Rótulos** | | | | | |
| **ET1-tf** | system, video retrieval, watermarking systems | models built, mr models, models | experimental research, theory, social interactions | highly effective, displacement sensors, effective bacterial adsorbents | research, velocity field, researchers |
| **ET2-tf** | system, hamiltonian systems, high capacity digital watermarking systems material | models built, mr models, models | experimental research, theory, social interactions | technical measurement, laser displacement, measured profiles | researchers, research, velocity field |
| **ET3-tf** | system, watermarking systems, video retrieval system | models built, mr models, models | experimental research, theory, social interactions | highly effective, highly effective bacterial adsorbents, laser scale | researchers, research, velocity field |
| **ET1-mdeg** | system, retrieval system, watermarking systems | models, model, models built | research, experimental research, research project | highly effective, measured, high shock | research, velocity field, researchers |

| | | | | | |
|---|---|---|---|---|---|
| **ET1-medeg** | system, retrieval system, watermarking systems | models, model, models built | research, experimental research, research project | highly effective, high shock, effective magnetic | research, velocity field, future empirical |
| **ET1-deg/tf** | reduced deterministic system avoids time consuming computations, high capacity digital watermarking systems material, traditional dct based watermarking systems | robust approximate tsk fuzzy model, traditional iteration modeling procedure, model named integral delay | experimental research project, social interactions, research question | direction high shock reliability test, highly effective bacterial adsorbents, highly effective paramagnetic complexes | structuring future empirical investigations, evaluating potential kbs applications, agile development approaches |
| **ET2-deg/tf** | retrieval system, hamiltonian systems, parallel systems | fuzzy modeling, modeling method, proposed modeling | experimental research, research project, social interactions | high degree, laser displacement, static laser | agile development, initiative development, development approaches |
| **ET3-deg/tf** | systems material, deterministic system, system avoids | iteration modeling, fuzzy modeling, modeling procedure | experimental research, research project, social interactions | direction high, high shock, highly effective | future empirical, structuring future, kbs applications |
| | | | | | |
| **Tópico** | learning, students, study, student, educational | large, efficient, computational, complexity, significantly | scheduling, resource, time, tasks, task | gene, biological, protein, http, genes | system, consistency, concurrent, transactions, distributed |
| **Rótulos** | | | | | |
| **ET1-tf** | reinforcement learning, significantly | public sentiment, time analysis, expressed public sentiment | periodic tasks, scheduling soft | availability, contact, http | distributed systems, systems requires, |

| | | | | | |
|---|---|---|---|---|---|
| | fewer, studying lifetime | | aperiodic tasks, soft aperiodic | | distributed systems requires |
| **ET2-tf** | studying lifetime, reinforcement learning, reinforcement learning approaches specifically designed | public sentiment, time analysis, expressed public sentiment | scheduling soft aperiodic tasks, periodic tasks, equivalent computational resources | availability, contact, http | distributed systems, distributed systems requires, systems requires |
| **ET3-tf** | reinforcement learning, studying lifetime, reinforcement learning approaches | public sentiment, time analysis, expressed public sentiment | scheduling soft aperiodic tasks, periodic tasks, aperiodic tasks | availability, contact, http | distributed systems, distributed systems requires, systems requires |
| **ET1-mdeg** | reinforcement learning, learning approach, learning approaches | time, public sentiment, costly dsms | periodic tasks, soft aperiodic, scheduling soft | contact, availability, http | system, distributed systems, systems requires |
| **ET1-medeg** | reinforcement learning, learning approach, learning approaches | time, public sentiment, costly dsms | periodic tasks, soft aperiodic, scheduling soft | contact, availability, genetic algorithm | distributed systems, systems requires, system |
| **ET1-deg/tf** | reinforcement learning approaches specifically designed, significantly fewer learning experiences, reinforcement | 3 times higher output rate compared, achieve efficient join window partitioning, executing costly dsms operators | linear dynamic texture analysis, modelling real p2p networks, real data tests illustrate | practical gene filtering approach based, general xml logical data model, short protein fragments derived | distributed monitoring tool called, tolerant distributed data base, tolerant distributed applications |

| | | | | | |
|---|---|---|---|---|---|
| | learning approach | | | | |
| **ET2-deg/tf** | reinforcement learning, learning approach, studying lifetime | making efficient, codewords costs, fast fading | scheduling approach, hard real, periodic tasks | protein data, protein similarity, biological role | tolerant distributed, distributed applications, distributed systems |
| **ET3-deg/tf** | learning approaches, reinforcement learning, learning experiences | achieve efficient, efficient join, executing costly | scheduling soft, aperiodic tasks, real data | practical gene, gene filtering, protein data | distributed monitoring, distributed data, tolerant distributed |
| | | | | | |
| **Tópico** | test, testing, tests, generation, generate | fuzzy, decision, making, uncertainty, theory | image, images, recognition, video, object | clustering, cluster, clusters, algorithm, hierarchical | international, papers, conference, proceedings, workshop |
| **Rótulos** | | | | | |
| **ET1-tf** | test cases, regression testing, generated test | decision making, multiple attribute decision, multiple attribute | optimal features, face recognition, features | number, genetic algorithms, application management | research problem, research findings, latest research |
| **ET2-tf** | existing regression testing methods generate test cases, generated test cases reveal, generated test cases uncover | multiple attribute decision making method, multiple attribute decision making problems, proposed | optimal features, features, images | clusters, number, genetic algorithms | research problem, research findings, latest research |
| **ET3-tf** | test cases, generated test cases, generated test | decision making, attribute decision making, decision making method | optimal features, features, features improves | clusters, number, genetic algorithms | research problem, research findings, latest research |

| | | | | | |
|---|---|---|---|---|---|
| **ET1-mdeg** | test cases, test, regression testing | decision making, attribute decision, multiple attribute | optimal features, features, face recognition | algorithm, genetic algorithm, number | research problem, research findings, latest research |
| **ET1-medeg** | test cases, regression testing, test | decision making, attribute decision, multiple attribute | optimal features, face recognition, features | genetic algorithm, genetic algorithms, application management | research problem, research findings, latest research |
| **ET1-deg/tf** | automated regression test generation regression testing involves testing, existing regression testing methods generate test cases, oriented automated test data generation assertions | multiple attribute decision making problems, multiple attribute decision making method, linear complexity encoding method | current image smoothing techniques, level view image retrieval, image retrieval work | based cluster application management system called appmanager, pocl story planning algorithm implements, cluster application management plays | traditional research problem, open research problem, latest research findings |
| **ET2-deg/tf** | regression testing, existing testing, testing methods | attribute values, encoding method, element method | image retrieval, visually similar, image filtering | cluster systems, existing algorithms, genetic algorithm | research problem, research findings, latest research |
| **ET3-deg/tf** | existing testing, generation assertions, regression testing | attribute decision, making problems, decision making | view image, image smoothing, current image | based cluster, cluster application, algorithm implements | research problem, research findings, latest research |
| | | | | | |
| **Tópico** | security, secure, | probability, distribution, model, markov, random | equations, numerical, equation, differential, method | paper, account, proposed, | security, privacy, secure, access, users |

| | | | | | |
|---|---|---|---|---|---|
| | scheme, key, schemes | | | different, approaches | |
| **Rótulos** | | | | | |
| **ET1-tf** | fusion scheme, feature fusion, feature fusion scheme | time, balanced random network model attracts considerable interest, network model | type equations, kdv type equations, kdv type | paper, propose, paper presents | users, heavy user request, heavy user |
| **ET2-tf** | fusion scheme, feature fusion scheme | balanced random network model attracts considerable interest, time, markov chain | type equations, kdv type equations, methods | paper, proposed, proposed network splitting algorithms | users, security, protection |
| **ET3-tf** | fusion scheme, feature fusion scheme | balanced random network model attracts considerable interest, random network model, time | type equations, kdv type equations, discontinuous galerkin method | paper, proposed, proposed network splitting | users, heavy user request, heavy user |
| **ET1-mdeg** | feature fusion, fusion scheme, feature fusion scheme | model, time, network model | method, kdv type, type equations | paper, propose, proposed | users, user request, heavy user |
| **ET1-medeg** | feature fusion, fusion scheme, feature fusion scheme | model, network model, random network | kdv type, type equations, method | paper, propose, network splitting | user request, users, heavy user |
| **ET1-deg/tf** | feature fusion scheme | balanced random network model attracts considerable interest, resultant equilibrium weight distribution, ingenious model structures | legendre spectral collocation methods, existing methods detect jumps, local discontinuous galerkin method | provide extensive experimental results, proposed network splitting algorithms, blind source separation approach | heavy user request traffic, security, protection |
| **ET2-deg/tf** | fusion scheme | kinetics model, ingenious model, model structures | type equations, segmentation methodology, boundary conditions | experimental results, results show, simulation results | security, protection, users |

| | | | | | |
|---|---|---|---|---|---|
| **ET3-deg/tf** | fusion scheme | balanced random, random network, model attracts | collocation methods, methods detect, existing methods | proposed network, separation approach, experimental results | heavy user, user request, users |
| | | | | | |
| **Tópico** | mean, resolution, satellite, correlation, sensing | robot, motion, control, robots, force | fault, failure, failures, tolerance, reliability | storage, system, memory, hardware, systems | control, controller, stability, systems, linear |
| **Rótulos** | | | | | |
| **ET1-tf** | remote sensing, spatial resolution, hyperspectral remote sensing | systems working, robotic systems working closely, working closely | error recovery, data errors, including unrecoverable ecc errors | operating system, performance, idisk systems | proposed, proposed controller, feedback control methodologies |
| **ET2-tf** | hyperspectral remote sensing, hyperspectral remote, remote sensing | robotic systems working closely, critical physical human robot interaction, control scheme | including unrecoverable ecc errors, time error recovery scheme, errors originated | idisk systems, performance, operating system | proposed controller, proposed, qft controller |
| **ET3-tf** | remote sensing, hyperspectral remote sensing, hyperspectral remote | robotic systems working closely, physical human robot, critical physical human robot interaction | error recovery, data errors, including unrecoverable ecc errors | idisk systems, performance, operating system | proposed controller, proposed, feedback control methodologies |
| **ET1-mdeg** | remote sensing, spatial resolution, hyperspectral remote | robot, systems working, physical human | error, error recovery, data errors | system, operating system, performance | proposed, gain control, feedback control |

| | | | | | |
|---|---|---|---|---|---|
| **ET1-medeg** | remote sensing, spatial resolution, hyperspectral remote | robot, systems working, physical human | error recovery, error, data errors | operating system, system, performance | proposed, gain control, feedback control |
| **ET1-deg/tf** | landsat multispectral scanner mss classifications, 5 high geometric resolution hrg-based, eucalyptus grandis plantation remote sensing | human finger strokes object surfaces, ridge walking motion planning algorithm, output feedback $h_\infty$ control | tmr redundant processor systems cache data errors read, level unequal error control codes, cpu control flow error | send high bit rate delay, high error rate compared, grained shared memory accesses | constant force feedback mechanism based, intelligent feedback control methodologies, linear variable gain amplifier |
| **ET2-deg/tf** | spatial resolution, coarse resolution, resolution data | human perception, represent human, human consciousness | tolerant systems, data errors, recovery procedures | shared memory, support systems, video system | based control, loop control, control scheme |
| **ET3-deg/tf** | resolution hrg-based, geometric resolution, landsat multispectral | $h_\infty$ control, physical human, robot interaction | tolerant systems, errors read, data errors | send high, high bit, high error | feedback mechanism, force feedback, linear gain |
| | | | | | |
| **Tópico** | program, code, language, programming, programs | role, dynamics, model, evolution, behavior | development, project, engineering, projects, process | language, text, word, words, english | mobile, devices, device, wireless, smart |
| **Rótulos** | | | | | |
| **ET1-tf** | programming language, programming | component models, standard java, java component | systems development, quality requirements, | visual languages, string languages, parsing visual | sensors, crowd computing, |

| | | | | | |
|---|---|---|---|---|---|
| | conventions, programming styles | | systems development process | | ubiquitous crowdsourcing |
| **ET2-tf** | programming language, programming conventions, programming styles | component models, standard java component models, mixture models | quality requirements, delivery processes, development project | visual languages, linguistic context, predefined linguistic context | sensors, crowd computing, ubiquitous crowdsourcing |
| **ET3-tf** | programming language, programming conventions, programming styles | component models, mixture models, java component models | quality requirements, systems development process, user quality requirements | visual languages, string languages, predefined linguistic context | sensors, crowd computing, ubiquitous crowdsourcing |
| **ET1-mdeg** | program, programming, language | model, component models, java component | systems development, quality requirements, development process | visual languages, visual language, parsing visual | sensors, wireless network, term wireless |
| **ET1-medeg** | program, programming, programming language | component models, model, java component | systems development, quality requirements, development process | visual languages, visual language, parsing visual | wireless network, sensors, term wireless |
| **ET1-deg/tf** | advanced language features, formal specification language, embed language abstractions | low complexity decoding scheme, evolutionary algorithm called biogeography, standard java component models | based systems development process meaningful user involvement, computer science education community unproductively assume, computer science students make productive | penn treebank corpus show, aware visual language editors, integrated language runtime | term wireless network usage, wireless lans, crowd computing |
| **ET2-deg/tf** | adaptable language, embed language, adapt language | component models, protein complexes, mixture models | based development, systems development, user quality | integrated language, language runtime, language syntax | wireless lans, crowd computing, ubiquitous crowdsourcing |

| | | | | | |
|---|---|---|---|---|---|
| **ET3-deg/tf** | adaptable language, embed language, adapt language | evolutionary algorithm, complexity decoding, low complexity | based development, process meaningful, unproductively assume | corpus show, treebank corpus, language editors | wireless network, term wireless, wireless lans |
| | | | | | |
| **Tópico** | neural, network, networks, learning, artificial | hardware, architecture, fpga, processor, system | services, service, architecture, web, oriented | detection, detect, detecting, false, monitoring | social, people, users, interaction, game |
| **Rótulos** | | | | | |
| **ET1-tf** | fuzzy model, 2 fuzzy, 2 fuzzy models | system, design model, design | service oriented architecture, service oriented, oriented architecture | time live, live streaming, time live streaming | twitter users, users post questions, surveyed twitter |
| **ET2-tf** | fuzzy model, 2 fuzzy models, fuzzy model leads | system, embedded system, design model | service oriented architecture, underlying service oriented architecture, service oriented | real, process execution time, practical engineering optimization problems involving real | twitter users post questions, contradicting users, twitter users |
| **ET3-tf** | fuzzy model, 2 fuzzy models, fuzzy model produced | system, embedded system, design model | service oriented architecture, service oriented, oriented architecture | time live, time live streaming, process execution time | twitter users, users post questions, twitter users post questions |
| **ET1-mdeg** | fuzzy model, model, 2 fuzzy | system, design, design model | service oriented, oriented architecture, service oriented architecture | live streaming, time live, time live streaming | twitter users, user, users post |
| **ET1-medeg** | fuzzy model, 2 fuzzy, model | system, design, design model | service oriented, oriented architecture, service oriented architecture | live streaming, time live, time live streaming | twitter users, users post, post questions |
| **ET1-deg/tf** | gradually constructs meaningful fuzzy partitions, | web service architecture providing qos management, fuzzy decision support | real web text data illustrate, based location dependent data services, class | time live streaming applications multipath streaming protocols, practical | twitter users post questions, ultimately increase user satisfaction, |

| | | | | | |
|---|---|---|---|---|---|
| | construct probabilistic fuzzy rule base, interpretable probabilistic fuzzy rule | system, large computer systems | support vector machines | engineering optimization problems involving real, time live streaming applications | helps people create meals |
| **ET2-deg/tf** | ranking model, iterative learning, model leads | computer systems, design method, oohdm design | service oriented, oriented architecture, based applications | time live, execution time, processing time | twitter users, collaborative cooking, contradicting users |
| **ET3-deg/tf** | fuzzy partitions, meaningful fuzzy, fuzzy rule | architecture providing, service architecture, support system | underlying service, real web, web text | involving real, time live, lowest false | users post, increase user, user satisfaction |
| | | | | | |
| **Tópico** | production, manufacturing, planning, demand, supply | surface, points, shape, geometric, surfaces | study, research, attention, literature, studies | channel, channels, interference, signal, error | retrieval, documents, search, document, text |
| **Rótulos** | | | | | |
| **ET1-tf** | optimization model, mathematic optimization, fuzzy time series | probabilistic points, tracing surfaces, ray tracing | research, experimental studies, studies | orthogonal frequency, frequency division multiplexing, frequency division | text documents, text density, document clustering |
| **ET2-tf** | ship unsinkability mathematic optimization model, advanced detonation wave tracking models, power supply | probabilistic points, points, ray tracing surfaces | research, studies, independently conducted controlled experimental studies | transmission radius, interference levels, misdetection errors | text documents, grouping documents, documents |

| | | | | | |
|---|---|---|---|---|---|
| **ET3-tf** | optimization model, mathematic optimization model, ship unsinkability mathematic optimization model | probabilistic points, points, ray tracing surfaces | research, experimental studies, studies | orthogonal frequency, frequency division multiplexing, frequency division | text documents, grouping documents, documents |
| **ET1-mdeg** | model, time series, mathematic optimization | tracing surfaces, ray tracing, points | research, experimental studies, studies | frequency division, division multiplexing, orthogonal frequency | search, text density, document clustering |
| **ET1-medeg** | model, time series, mathematic optimization | tracing surfaces, ray tracing, ray tracing surfaces | experimental studies, research, studies | frequency division, division multiplexing, orthogonal frequency | text density, document clustering, search |
| **ET1-deg/tf** | minimum cost flow network problems, advanced detonation wave tracking models, incredible shrinking bug separation model | dimensional linear subspace computed, free surface flow context, ray tracing surfaces | independently conducted controlled experimental studies, experimental studies software review, experiment results show | orthogonal frequency division multiplexing systems, orthogonal frequency division multiplexing, iterated lp relaxation framework | selectively diversifying web search results search result diversification, researchers utilized statistical regression, combine information filtering techniques |
| **ET2-deg/tf** | optimization model, model reliability, validate model | tracing surfaces, bidimensional maxwell, 3d cues | experimental studies, experiment results, results show | frequency domain, error protection, unequal error | search results, chinese search, search relevance |
| **ET3-deg/tf** | tracking models, minimum cost, cost flow | dimensional linear, surface flow, free surface | studies software, experimental studies, results show | iterated lp, orthogonal frequency, channel modeling | web search, search result, search results |

# APÊNDICE E – RESULTADOS SIGIR

| **Tópico** | Data, visual, system, information, graphic, representation | Approach, classification, main, text, algorithm | Large, digital, provide, databases, Language | Based, similar, probability, document, very | Summarization, high, generate, selection, construct |
|---|---|---|---|---|---|
| **Rótulos** | | | | | |
| **ET1-tf** | Visualizations, systems, representation | Classification, textual, linear | Databases, management systems, statistical | Topic model, similarity based, probability | Summary, summarization, sentence selection |
| **ET2-tf** | Visualizations, systems, model | Classification, textual, algorithm | Databases, management systems, statistical | Topic model, probability, similarity based | Summary, sentence selection, construct summaries |
| **ET3-tf** | Visualizations, model, information | Classification, textual, algorithm | Databases, management systems, statistical | similarity based probability, model, latent semantic indexing | Summary, summarization, construct summaries |
| **ET1-mdeg** | Visualizations, information, information representation | Text classification, textual, algorithm | Databases, management systems, database systems | Topic model, probability, latent semantic indexing | summarization, summary, summarization research |
| **ET1-medeg** | Visualizations, information representation, graphic model | Classification, classification algorithm, textual data | Databases, management systems, database systems | Latent semantic indexing, topic model, probability | summarization, summary, summarization research |
| **ET1-deg/tf** | Data flow, visualizations, data analysis | Textual data used, supervised approaches, algorithm | Databases, become very large, information systems | Latent semantic indexing, similarity based probability model, statistical technique | Highly-condensed, extractive summaries, sentence selection |
| **ET2-deg/tf** | Data flow, analysis, data visualization | Textual data, classification algorithm, supervised approaches | Databases, become very large, information systems | statistical technique, Latent semantic indexing, similarity based probability model | Highly-condensed, extractive summaries, statistical approach |
| **ET3-deg/tf** | Graphic usage, data analysis, data | Textual data used, supervised approaches include, main classification algorithm | Databases, base management system, become very large | similarity based probability model, statistical technique, Latent semantic indexing | Highly-condensed, sentence selection, summarizing text documents |

| **Tópico** | Parallel, language, cross, comparison, translation | Build, distributed, problem, efficient, retrieval | English, methods, Japanese, language, query | Boolean, belief, use, extend, operator | Solution, probabilistic, models, selection, paper |
|---|---|---|---|---|---|
| **Rótulos** | | | | | |
| **ET1-tf** | Translation, query translation, information | Distributed, efficient algorithms, information | Japanese, retrieval, indexing | Revision operator, Boolean operator, belief revision | Information retrieval, probabilistic model, learning |
| **ET2-tf** | query translation, information access, translation | Distributed, efficient algorithms, information | Japanese text, retrieval, indexing | Revision operator, Boolean operator, extended | Information retrieval, probabilistic model, adaptive solution |
| **ET3-tf** | query translation, information access, translation | Distributed, efficient algorithms, information | Japanese text, information retrieval, indexing | Revision operator, extended, model | Information retrieval, probabilistic model, adaptive solution |
| **ET1-mdeg** | Cross-language, parallel texts, translation | Distributed, information retrieval, efficient algorithms | Japanese text, information retrieval, Chinese | Boolean operator, revision operator, extended | Probabilistic model, language, information retrieval |
| **ET1-medeg** | Translation, cross-language, information access | Distributed information retrieval, efficient algorithms, fusion problem | Japanese text, Chinese information retrieval, indexing | Boolean operator, revision operator extended Boolean model | Probabilistic model, language, information retrieval |
| **ET1-deg/tf** | Cross-language information retrieval, query translation methods, information access across | Probabilistic solution, efficient distributed algorithms, build inverted files | English Japanese texts, comparing representations, query translation methods | Belief revision operator, document ranking, computational | Adaptive filtering agent, agent based, learning model |
| **ET2-deg/tf** | Cross-language information retrieval, translation methods, information access | Probabilistic solution, efficient distributed algorithms, build inverted files | English Japanese texts, comparing representations, query translation methods | Belief revision operator, document ranking, computational | Adaptive filtering agent, agent based, learning model |
| **ET3-deg/tf** | Cross-language information retrieval, query translation methods, information access across | efficient distributed algorithms, build inverted files, Probabilistic solution | English Japanese texts, comparing representations, query translation methods | Belief revision operator, document ranking, extended Boolean model | Interaction modelling, agent based, information retrieval |

| **Tópico** | Content, network, link, work, present | Evaluation, give, batch, results, measure | User, increase, extensive, search, query | Using, class, hypertext, compare, hyperlink | Technique, similar, filter, filtering, general |
|---|---|---|---|---|---|
| **Rótulos** | | | | | |
| **ET1-tf** | Link, cross reference, information retrieval | Evaluation, novel method, improvement | Search, engines, terms | Hypertext, hyperlink, compare class | Filtering, filter, users |
| **ET2-tf** | Link, cross reference, information retrieval | Evaluation, novel method, improvement | Web Search, search engines, terms | Hypertext, hyperlink, compare class | Filtering, filter, users |
| **ET3-tf** | Link, cross reference, information retrieval | Evaluation, novel method, results show | Web Search, engines, interactive search | Hypertext, information retrieval, compare class | Filtering, filter, users |
| **ET1-mdeg** | Link, web, information retrieval | Evaluation, large-scale, novel method | Search engines, web search, search | Hypertext, information retrieval, hyperlink | Filtering, user profiles, filter |
| **ET1-medeg** | Link, information retrieval, cross reference | User evaluation, evaluation methods, novel method | Search engines, web search, search | Hypertext, hyperlink, information retrieval | Filtering, various techniques, document similarity |
| **ET1-deg/tf** | Information retrieval model, reference collection extracted, information derived | Large-scale evaluation, evaluation, information retrieval system | search term logs, Interactive web search, contain representative | Hypertext categorization using, high demand, recently proposed | Information filtering, user profile from, calculate similarity between |
| **ET2-deg/tf** | Information retrieval model, reference collection extracted, information derived | Large-scale evaluation, evaluation, information retrieval system | Interactive web search, search term logs, contain representative search term | Hypertext categorization using, high demand, recently proposed | Information filtering, user profile from, calculate similarity between |
| **ET3-deg/tf** | Information retrieval model, reference collection extracted, evidencial information | First Large-scale evaluation, evaluation, information retrieval system | Interactive web search, search term logs, contain representative | Assisted search usually, recently proposed, Hypertext categorization using | collaborative filtering, user profile from, using document content |

| **Tópico** | Detection, effect, near, news, tracking | Clustering, cluster, based, algorithm | Knowledge, information, usually, base, present | Word, effect, SDR, spoken, sets | Annotation, call, answers, identify, question |
|---|---|---|---|---|---|
| **Rótulos** | | | | | |
| **ET1-tf** | TDT, corpora, news | Cluster, model based, distribution | Knowledge, present, base | Item set, SDR, vocabulary word | QA, question, answer |
| **ET2-tf** | TDT, news source, document | Cluster, distribution, text features | Knowledge base, experiments, validates several | Item set, document retrieval, information retrieval | QA, answering, identify potential |
| **ET3-tf** | TDT, news source, document | Clustering algorithms, model based, joint distribution | Information retrieval systems, knowledge, paper presents | Item set effects, vocabulary word, document retrieval | QA, question, answer |
| **ET1-mdeg** | TDT, detection, tracking | Clustering, clusters, distribution | Knowledge, present, information retrieval | SDR, item sets, information retrieval | QA, answering, question |
| **ET1-medeg** | TDT, topic detection, tracking | Clustering algorithms, clusters, performance | Information retrieval, knowledge, paper presents | SDR, document retrieval, information retrieval | QA answering, answers |
| **ET1-deg/tf** | Broadcast news stories, two questions, background collection | Clustering, clusters, performance compared | Design experiments capable, knowledge base, information retrieval | Item sets, vocabulary words, recognition system | Annotation identify potential, seeking questions posed, legitimate differences |
| **ET2-deg/tf** | Topic detection, background collection, news stories | Clustering, clusters, algorithms | Design experiments, knowledge base, paper presents | Item sets effects, vocabulary words, speech recognition system | Annotation identify potential, seeking questions posed, legitimate differences |
| **ET3-deg/tf** | Topic detection, text improve performance, features | Clustering, clusters, model based | Design experiments, information retrieval systems, paper presents | Item sets effetcs, vocabulary words, speech recognition system | Question answering system, Annotation identify potential |

| **Tópico** | File, index, signature, access, hash | Association, image, digital, domain, process | Indexing, weighting, theory, combination, component | Explore, inverted, structure, size, fast | Feedback, ranking, human, assign, relevance |
|---|---|---|---|---|---|
| **Rótulos** | | | | | |
| **ET1-tf** | File, access methods, signatures | Image, requirement, user | Term, importance, weighting | Inverted lists, fast, file structure | Relevance, human feedback, feedback |
| **ET2-tf** | Signature files, access methods, signatures | Image, user model, form | Weighting, term importance, component | Inverted lists, fast, file structure | Relevance, human feedback, ranking documents |
| **ET3-tf** | Signature files, access methods, signatures | Image form, images, user model | Term importance, weighting term, document components | Inverted lists, fast, inverted files | Relevance, human feedback, ranking documents |
| **ET1-mdeg** | File, signatures, access methods | Image model, image, user model | Term, importance, component | Fast, inverted lists, inverted files | Relevance, human, feedback |
| **ET1-medeg** | Signature files, file, signatures | Image form, user requirements, user model | Term importance, document components, weighting | Inverted lists. Inverted files, fast | Relevance, feedback, human |
| **ET1-deg/tf** | Dynamic signature technique, retrieval system, access methods | User requirements, user model, image model | Combination match, term importance, unified interpretation | Disk based inverted file, many times faster, integer compression schemes | Ranking documents between, relevance feedback techniques, unknown relevance |
| **ET2-deg/tf** | Dynamic signature technique, retrieval system, access methods | Image model, user model, requirements | Principle integrates, combination match, term importance | Disk based inverted file, integer compression schemes, fast query evaluation | Ranking documents between, relevance feedback techniques, unknown relevance |
| **ET3-deg/tf** | New signature technique, retrieval system, access methods | Independent image retrieval, user requirements, image form | Term importance, term combination, term importance | Disk based inverted file, integer compression schemes, fast query evaluation | Ranking documents between, relevance feedback techniques, high ranked documents |

| **Tópico** | Compression, full , being, new, approximation | Model, semantic, develop, schema, vector | Adaptive, framework, implementation, describe, composed |
|---|---|---|---|
| **Rótulos** | | | |
| **ET1-tf** | Compression technique, new methods, full text | Vector, schema, semantic | Implementation, application, describe two |
| **ET2-tf** | Compression technique, new methods, full text | Vector space model, logical model, semantic model | Describe two, different implementation, compare framework |
| **ET3-tf** | Compression technique, full text document, new methods | Vector space model, logical model, semantic model | Describe two, different implementation, compare framework |
| **ET1-mdeg** | Compression technique, compression, new methods | Vector space model, semantic model, schema | Implementation, approach, compares two |
| **ET1-medeg** | Compression, compression technique, new methods | Vector space model, schema, semantic model | Implementation, two applications, compare |
| **ET1-deg/tf** | Compression technique, full text, new methods | Vector space model, logical model, semantic model | Paper compares two different, framework described here, first approach |
| **ET2-deg/tf** | Compression technique, full text, memory document retrieval | Relational model, bibliographic system schema, vector space model | Paper compares two different, framework described here, first approach |
| **ET3-deg/tf** | Posting Compression technique, dynamic full text document, compressing such matrices | Vector space model, semantic binary relationship model, experimental semantic model user interface | Paper compares two different, framework described here, first approach |

# APÊNDICE F – RESULTADOS SBBD

| Ano | 1999 | | | | |
|---|---|---|---|---|---|
| **Tópico** | Dimension, migration, execution, network, multi | Local, database, transaction, failure, distributed | Essential, driven, active, event, knowledge | Dvr, module, metadata, dimensional, sneps | Index, conventional, base, example, show |
| **Rótulos** | | | | | |
| **ET1-tf** | database migration, scheduling method, wan environments | integrated transaction, session facilities, supporting | user interfaces, integrity constraints, data-driven active | no conventional data, data warehouse, series | Base, using frameworks, sharing |
| **ET2-tf** | database migration, database engine, wan environments | integrated transaction, fuzzy data, supporting | user interfaces, derived data, data-driven active | Data warehouse, ambients, classification | Base, modules, sharing |
| **ET3-tf** | database migration, database engine, object oriented | integrated transaction, session facilities, data | user interfaces, databases, data-driven active | Data warehouse, classification, no conventional data | Base, using frameworks, systems |
| **ET1-mdeg** | database migration, scheduling method, wan environments | integrated transaction, session facilities, supporting | Data-driven active, database management, knowledge | no conventional data, data warehouse, series | Base, using frameworks, sharing |
| **ET1-medeg** | database migration, scheduling method, wan environments | integrated transaction, session facilities, supporting | user interfaces, derived data, data-driven active | statics abstracts, obtainment, no conventional data | Base, using frameworks, sharing |
| **ET1-deg/tf** | object oriented database engine, scheduling method, wan environments | Fuzzy data, bases, integrated transaction | knowledge base management, ontology-aware database management, towards | object relational technology, mining, statics abstracts | Heterogeneous data using framework, integration, bases |
| **ET2-deg/tf** | reflective persistence middleware, scheduling method, database engine | Uncertainty, session facilities, fuzzy data | Knowledge base management, integrity constraints, database | object relational technology, mining, data warehouse | Heterogeneous data using framework, modules, bases |
| **ET3-deg/tf** | Reflective persistence middleware, scheduling method, database engine | Uncertainty, session, facilities, fuzzy data | DBMS, KBMS, derived data | object relational technology, mining, data warehouse | Heterogeneous data using framework, modules, bases |

| Ano | 1999 | | 2000 | | |
|---|---|---|---|---|---|
| **Tópico** | Query, persistence, action, fuzzy, inheritance | Temporal, data, object, version, schema | Temporal, program, constraint, time, show | Client, broadcast, multimedia, insert, run | Warehouse, spatial, factor, system, implement |
| **Rótulos** | | | | | |
| **ET1-tf** | query, relational data, graphic interface | schema versioning, indexation, objects | temporal database, conventional database, concurrency | Multimedia metadata, modeling, insertion | Data warehouse systems, implementing, metadata database |
| **ET2-tf** | query, relational data, inheritance | schema versioning, objects, techniques | temporal database, conventional database, approach | Multimedia metadata, insertion, corporations | Data warehouse systems, database management, implementing |
| **ET3-tf** | query, inheritance, relational data | Schema versioning, techniques, indexation | Temporal database, approach, concurrency | Multimedia metadata, corporations, modeling | Data warehouse systems, database management, metadata database |
| **ET1-mdeg** | query, relational data, graphic interface | schema versioning, indexation, objects | temporal database, conventional database, concurrency | Multimedia metadata, modeling, insertion | Data warehouse systems, implementing, metadata database |
| **ET1-medeg** | query, relational data, graphic interface | schema versioning, indexation, objects | temporal database, conventional database, concurrency | Multimedia metadata, modeling, insertion | Data warehouse systems, implementing, metadata database |
| **ET1-deg/tf** | graphical notebook, interaction metaphor, querying databases | an experimental evaluation, oriented data, temporal databases supporting schema versioning | Data definition management, broadcast environments, data model | Knowledge discovery, methodology targeted, object-oriented data mart | Heterogeneous database management, environmental information, architecture |
| **ET2-deg/tf** | graphical notebook, querying databases, relational data | An experimental evaluation, queries, oriented data | Data definition management, temporal serialization graph testing, top | Knowledge discovery, object-oriented data mart, data warehouse | Heterogeneous database management, architecture, data warehouse systems |
| **ET3-deg/tf** | graphical notebook, querying databases, relational data | An experimental evaluation, queries, oriented data | Data definition management, temporal serialization graph testing, top | Knowledge discovery, object-oriented data mart, data warehouse | Heterogeneous database management, architecture, data warehouse systems |

| Ano | 2000 | | | 2001 | |
|---|---|---|---|---|---|
| **Tópico** | Interface, visual, email, not, large | Parallel, intersect, join, approximation, affect | Constraint, dimension, formula, past, attribute | Join, table, index, integration, web | r-tree, temporal, semistructured, approach, evolution |
| **Rótulos** | | | | | |
| **ET1-tf** | Visual query interface, data, large volumes | Affect, performance, processing | Temporal databases, genetic programming, databases | Data integration, web, agents | Temporal management, approach, evolution |
| **ET2-tf** | Visual query interface, large volumes, mail-by-example | Affect, factors, processing | Temporal databases, mining, genetic programming | Data integration, agents, mediators | Temporal management, schemes, evolution |
| **ET3-tf** | Visual query interface, mail-by-example, data | Affect, factors, performance | Databases, management systems, statistical | Data integration, web, mediators | Temporal management, schemes, approach |
| **ET1-mdeg** | Visual query interface, data, large volumes | Affect, performance, processing | Temporal databases, genetic programming, databases | Data integration, web, agents | Temporal management, approach, evolution |
| **ET1-medeg** | Visual query interface, data, large volumes | Affect, performance, processing | Temporal databases, genetic programming, databases | Data integration, web, agents | Temporal management, approach, evolution |
| **ET1-deg/tf** | Geographic information systems, geovisual interface, large volumes | Spatial joins, polyline joins, processing | Dynamic intregrity constraints, mining temporal constraints, genetic programming | Organize electronic documents, metadata approach, generation | Semi-structured data, relational data, storage |
| **ET2-deg/tf** | Geographic information systems, geovisual interface, visual query interface | Spatial joins, polyline joins, raster approximation | Dynamic intregity constraints, past-directed evaluation, mining temporal constraints | Organize electronic documents, metadata approach, collections | Semi-structured data, versions, storage |
| **ET3-deg/tf** | Geographic information systems, geovisual interface, visual query interface | Spatial joins, polyline joins, raster approximation | Dynamic intregity constraints, past-directed evaluation, mining temporal constraints | Organize electronic documents, metadata approach, collections | Semi-structured data, versions, storage |

| Ano | 2001 | | | | |
|---|---|---|---|---|---|
| **Tópico** | Image, maximal, itemset, visual, decision | Generation, tthesaurus, juridical, retrieval, tool | Medical, gis, evolution, collection, spatial | Atribute, file, olap, dimension, parallelism | Xml, management, data, mining, similar |
| **Rótulos** | | | | | |
| **ET1-tf** | Visualizations, systems, representation | Information retrieval, tool, linear | Spatiotemporal database, documenting, extending | Partitioned parallelism, applying, olap paradigm | Xml, constraints, data management |
| **ET2-tf** | Visualizations, systems, model | Information retrieval, classification, algorithm, tool | Spatiotemporal database, extending, model | Partitioned parallelism, olap paradigm, framework | Xml, data management, ugly |
| **ET3-tf** | Visualizations, model, information | Information retrieval, classification, algorithm, tool | Spatiotemporal database, model, documenting | Partitioned parallelism, framework, applying | Xml, constraints, ugly |
| **ET1-mdeg** | Visualizations, information, information representation | Information retrieval, classification, algorithm, tool | Spatiotemporal database, documenting, extending | Partitioned parallelism, applying, olap paradigm | Xml, constraints, data management |
| **ET1-medeg** | Visualizations, information representation, graphic model | Information retrieval, classification, algorithm, tool | Spatiotemporal database, documenting, extending | Partitioned parallelism, applying, olap paradigm | Xml, constraints, data management |
| **ET1-deg/tf** | Data flow, visualizations, data analysis | collections, personal names searching; flexible approximate | Ensure topological space constraints, opengis, documenting | Optimizer generator framework, incorporating deviation-detection functionality, olap paradigm | Personal names searching, flexible approximate, tool |
| **ET2-deg/tf** | Data flow, analysis, data visualization | collections, personal names searching; flexible approximate | Ensure topological space constraints, opengis, spatiotemporal database | Optimizer generator framework, incorporating deviation-detection functionality, parallelism | Personal names searching, flexible approximate, data integration |
| **ET3-deg/tf** | Graphic usage, data analysis, data | collections, personal names searching; flexible approximate | Ensure topological space constraints, opengis, spatiotemporal database | Optimizer generator framework, incorporating deviation-detection functionality, parallelism | Personal names searching, flexible approximate, data integration |

| Ano | 2002 | | | | |
|---|---|---|---|---|---|
| **Tópico** | Encrypted, protocol, secure, data, extraction | Database, mobile, sharing, approach, object | Schema, xml, documents, dbms, ontology | Processing, query, strategy, join, dig | Language, cross, available, keyword, base |
| **Rótulos** | | | | | |
| **ET1-tf** | Protocol, encrypted data, database | Amdb, approach, object relational | Xml, schemas, relational databases | Statistics, queries, mm service | Available data, keywords, querying bases |
| **ET2-tf** | Protocol, encrypted data, database | Ambd, approach, detection | Xml documents, schemas, relational databases | Statistics, mm service, parallel joins | Available data, querying bases, cross-language |
| **ET3-tf** | Protocol, encrypted data, database | Amdb, detection, environments | Xml documents, schemas, relational databases | Statistics, queries, parallel joins | Available data, keywords, cross-language |
| **ET1-mdeg** | Protocol, database, encrypted data, | Amdb, approach, object relational | Xml doduments, xml, schemas | Statistics, queries, mm service | Available data, keywords, querying bases |
| **ET1-medeg** | encrypted data, protocol, database | Amdb, databases, object relational | Xml documents, schemas, relational databases | Statistics, queries, mm service | Available data, cross-language, keywords |
| **ET1-deg/tf** | data exchange protocol, relational algebra operations, strong key management | Sharing mobile databases, mobile computing, support | works store XML documents, object-based representation, semantic xml-schemas | Competitive online comparison, distributed processing, queries | Personalized keyword search, partial-order preferences, querying bases |
| **ET2-deg/tf** | data exchange protocol, relational algebra operations, strong key management | Sharing mobile databases, mobile computing, format-independent | works store XML documents, object-based representation, semantic xml-schemas | Competitive online comparison, distributed processing, parallel joins | Personalized keyword search, partial-oreder preferences, web using keywords |
| **ET3-deg/tf** | Secure database, analytical-based decision processes, knowledge management | Sharing mobile databases, mobile computing, format-independent | works store XML documents, object-based representation, semantic xml-schemas | Competitive online comparison, distributed processing, parallel joins | Personalized keyword search, partial-oreder preferences, web using keywords |

| Ano | 2002 | 2003 | | | |
|---|---|---|---|---|---|
| **Tópico** | Decision, association, rules, support, mining | Xml, metadata, proposal, repositories, management | Integration, provenance, systems, based, ontologies | Distributed, workflow, fragmentation, show, execution | Mining, clustering, tool, neighbor preparation |
| **Rótulos** | | | | | |
| **ET1-tf** | Association, mining, rules | Xml views, metadata, management | Data provenance, integration systems, queries | Proposed algorithm, distributed databases, environments | Mining, clustering, data preparation |
| **ET2-tf** | Association, mining algorithm, decision process | Xml views, metadata, proposal | Data provenance, integration systems, data | Proposed algorithm, distributed databases, workflow execution | Mining, clustering algorithm, data preparation |
| **ET3-tf** | Association, mining algorithm, decision process | Xml views, proposal, management | Data provenance, queries, domain ontologies | Propsed algorithm, distributed databases, workflow execution | Mining, nearest neighbor, data preparation |
| **ET1-mdeg** | Association, mining, rules | Xml views, metadata, management | Data provenance, integration systems, queries | Propsed algorithm, distributed databases, environments | Mining, clustering, data preparation |
| **ET1-medeg** | Association, mining algorithm, decision process | Xml views, metadata, management | Data provenance, integration systems, domain ontologies | Propsed algorithm, distributed databases, workflow execution | Mining, clustering algorithm, data preparation |
| **ET1-deg/tf** | Reliable models, dynamic databases, determine rules | Relational databases, xml standarts, uxquery | Xml-based data integration systems, generation mediation queries, provenance | Data modification language, temporal schema versioning, support | Data mining, indexing metrics, nearest neighbor method |
| **ET2-deg/tf** | Reliable models, dynamic databases, incremental algorithm | Relational databases, xml standarts, mof repositories | Xml-based data integration systems, generation mediation queries, domain ontologies | Data modification language, temporal schema versioning, relational databases | Data mining, indexing metrics, engine effectiveness |
| **ET3-deg/tf** | Reliable models, dynamic databases, incremental algorithm | Relational databases, xml standarts, mof repositories | Xml-based data integration systems, generation mediation queries, domain ontologies | Data modification language, temporal schema versioning, relational databases | Semantic web, indexing metrics, engine effectiveness |

| Ano | 2003 | | 2004 | | |
|---|---|---|---|---|---|
| **Tópico** | Detection, summaries, warehouse, source, mart | Performance, query, evaluation, search, similarity | Visual, mining, approach, first, based | Cluster, query, olap, process, dynamic, heterogeneous | Warehouse, metamodel, gdw, aquaware, quality |
| **Rótulos** | | | | | |
| **ET1-tf** | Source data, summaries, detection | Performance, query, evaluation | Approach, mining, visual analysis | Olap, Database cluster, query | Metamodels, aquaware, data warehousing |
| **ET2-tf** | Souce data, warehouses, data mart | Query performance, Performance, evaluation | Approach, mining, selection | Database cluster, query processing, data sources | Metamodels, data warehousing, gdw |
| **ET3-tf** | Souce data, summaries, data mart | Query performance, Performance, evaluation | Approach, visual analysis, selection | Database cluster, data sources, query processing | Metamodels, gdw, aquaware |
| **ET1-mdeg** | Source data, summaries, detection | Performance, query, evaluation | Mining, approach, visual analysis | Olap, Database cluster, query processing | Metamodels, aquaware, data warehousing |
| **ET1-medeg** | Source data, warehouses, data mart | Query Performance, evaluation, similarity queries | Mining, approach, visual analysis | Database cluster, query processing, data sources | Metamodels, data warehousing, data quality |
| **ET1-deg/tf** | Evaluating warehouses, improved approache, cluster | Quality evaluation, expensive predicates, performance | First-order temporal pattern mining, apriori-based approach, feature selection | Integrating heterogeneous data sources, dynamic environment, olap query processing | Geographical integration based, providing multidimensional, data warehousing |
| **ET2-deg/tf** | Evaluating warehouse, improved approach, data warehousing etlm process | Quality evaluation, expensive predicates, semantic query processing strategy | First-order temporal pattern mining, apriori-based approach, data mining processes | Integrating heterogeneous data sources, dynamic environment, adaptive virtual partitioning | Geographical integration based, providing multidimensional, data quality support environment |
| **ET3-deg/tf** | Evaluating warehouse, improved approach, data warehousing etlm process | Quality evaluation, expensive predicates, semantic query processing strategy | First-order temporal pattern mining, apriori-based approach, data mining processes | Integrating heterogeneous data sources, dynamic environment, adaptive virtual partitioning | Geographical integration based, providing multidimensional, data quality support environment |

| Ano | 2004 | | | | |
|---|---|---|---|---|---|
| **Tópico** | Data, distributed, parallel, environment, integrating | Terms, collaborative, model, lock, vector | Navigation, persist, temporal, version, store | optimization, web, case, engine, computation | Computing, schema, xml, basis, documents |
| **Rótulos** | | | | | |
| **ET1-tf** | Density-data, dbm-tree, metric | Terms, vector, model | Coding, framepersist, paths | Computations, case, web search engines | Xml, computing, schemas |
| **ET2-tf** | Density-data, parallel queries, integrating heterogeneous | Terms, model, lock | Temporal data, coding navigation framepersist | Computations, web search engines, optimization | Computing, xml, dependency basis |
| **ET3-tf** | Density-data, parallel queries, integrating heterogeneous | terms, lock, vector | Temporal data, coding navigation framepersist | Computations, web service, optimization | Computing, schemas, dependency basis |
| **ET1-mdeg** | Density-data, parallel queries, integrating heterogeneous | Terms, vector, model | Coding, framepersist, paths | Computations, case, web search engines | Xml, computing, schemas |
| **ET1-medeg** | Density-data, parallel queries, integrating heterogeneous | Terms, vector, model | Temporal data, coding navigation framepersist | Computations, web search engines, web service | Xml, computing, schemas |
| **ET1-deg/tf** | Dynamic metric access method sensitive, metric trees, achieve | Natively stored xml documents, collaborative processing, terms | Persistent object stores, mobile service applications, coding | Towards cost-based optimization, data-intensive web, web search engines | Xml schemas, relational schemas, dependency basis |
| **ET2-deg/tf** | Dynamic metric access method sensitive, metric trees, twisting | Natively stored xml documents, collaborative processing, dependence | Persistent object stores, mobile service applications, object persistence framework | Towards cost-based optimization, data-intensive web, optimizing ranking calculation | Xml schemas, relational schemas, nested list attributes |
| **ET3-deg/tf** | Dynamic metric access method sensitive, metric trees, twisting | Natively stored xml documents, collaborative processing, dependence | Persistent object stores, mobile service applications, object persistence framework | Towards cost-based optimization, data-intensive web, optimizing ranking calculation | Xml schemas, relational schemas, nested list attributes |

| Ano | 2005 | | | | |
|---|---|---|---|---|---|
| **Tópico** | Digital, web, semantic, self, applications | Rule, select, classification, tvcl, constraint | Documents, deweyids, schema, control, xml | Sql, objects, querying, complex, transformation | Genetic, similarity, author, library, ecologic |
| **Rótulos** | | | | | |
| **ET1-tf** | Web, searching, knowledge | Tvcl, rule selection, classification | Key, deweyids, xml documents | Sql, querying, transformation | Authorship, similarity queries, identification |
| **ET2-tf** | Web, knowledge base, applications | Tvcl, rule selection, classification | Key, xml documents, management | Sql, transformation, objects | Authorship, similarity queries, removal |
| **ET3-tf** | Web, searching, applications | Tvcl. rule selection, classification | Key, management, deweyids | Sql, objects, querying | Authorship, digital libraries, identification |
| **ET1-mdeg** | Web, searching, knowledge base | Tvcl, classification, rule selection | Xml documents, deweyids, integrity | Sql, querying, transformation | Authorship, similarity queries, identification |
| **ET1-medeg** | Web, searching, knowledge base | Tvcl, classification, rule selection | Xml documents, deweyids, integrity | Sql, querying, transformation | Authorship, similarity queries, digital libraries |
| **ET1-deg/tf** | Self describing components, geographic knowledge base, web | Temporal versioned constraint language, global self-tuning architecture, searching | Xml documents, fine-grained management, key | Complex objects, extending relational algebra, querying | Bibliographic objects, genetic algorithms, ambiguities |
| **ET2-deg/tf** | Self describing components, geographic knowledge base, semantic web applications | Temporal versioned constraint language, global self-tuning architecture, tvcl | Xml documents, fine-grained management, domain integrity constraint | Complex objects, extending relational algebra, one-to-many data transformations | Bibliographic objects, genetic algorithms, approximate similarity |
| **ET3-deg/tf** | Self describing components, geographic knowledge base, semantic web applications | Temporal versioned constraint language, global self-tuning architecture, tvcl | Xml documents, fine-grained management, domain integrity constraint | Complex objects, extending relational algebra, one-to-many data transformations | Bibliographic objects, genetic algorithms, approximate similarity |

| Ano | 2005 | | 2006 | | |
|---|---|---|---|---|---|
| **Tópico** | Peer, learning, p2p, rosa, data | Algorithm, pattern, mining, time, stream | Systems, mining, ontologies, fuzzy, pattern | Versioned, voql, language, query, versions | Rdbms, web, hmm, faqs, metadata |
| **Rótulos** | | | | | |
| **ET1-tf** | Data, p2p, mobile devices | Generation, classification, mining | Semantically, systems, data mining | versions, query language, voql | Rdbms, extracting, web faqs |
| **ET2-tf** | Data, peer-to-peer, compression | Data streams, classification, regular expression | Semantically, data mining, fuzzy logic | Complexity, database, voql | Rdbms, web faqs, digital libraries |
| **ET3-tf** | Data, mobile devices, peer-to-peer | Data streams, regular expression, mining | Semantically, data mining, fuzzy logic | Complexity, voql, query language | Rdbms, digital libraries, extracting data |
| **ET1-mdeg** | Data, compression, mobile devices | Generation, mining, classification | data mining , Semantically, systems | versions, query language, voql | Rdbms, extracting data, web faqs |
| **ET1-medeg** | Peer-to-peer, p2p, mobile devices, | Data streams, classification, mining | data mining, Semantically, fuzzy logic | versions, voql, query language | Rdbms, extracting data, web faqs |
| **ET1-deg/tf** | Biodiversity case, peer-to-peer databases, efficient architecture | Cost-sensitive associative classification, selection techniques, first-order temporal pattern mining | Fuzzy ontologies, automatic inconsistency, data mining | Versioned object oriented database, process pipeline scheduling, queries | Searching useful information, componentized digital libraries, web faqs |
| **ET2-deg/tf** | Biodiversity case, peer-to-peer databases, limited computing resources | Cost-sensitive associative classification, selection techniques, regular expression constraints | Fuzzy ontologies, automatic inconsistency, case-based reasoning systems | Versioned object oriented database, process pipeline scheduling, web | Searching useful information, componentized digital libraries, workflow support |
| **ET3-deg/tf** | Biodiversity case, peer-to-peer databases, limited computing resources | Cost-sensitive associative classification, selection techniques, regular expression constraints | Fuzzy ontologies, automatic inconsistency, case-based reasoning systems | Versioned object oriented database, process pipeline scheduling, web | Searching useful information, componentized digital libraries, workflow support |

| Ano | 2006 | | | 2007 | |
|---|---|---|---|---|---|
| **Tópico** | Dimension, model, object, warehouse, data | Active, views, xml, generation, documents | Object, algorithms, scale, support, large | Collections, libraries, web, computation, digital | Ajax, replix, join, manet, mobile |
| **Rótulos** | | | | | |
| **ET1-tf** | Application, data warehouses, objects | Generation, xml data, query | Algorithms, scale, ontologies | Web collections, digital libraries, clustering | Ajax, mdbc, mobile databases |
| **ET2-tf** | Application, dimensional modeling, data warehouses | Generation, xml data, query aware | Algorithms, ontologies, object reporitories | Web collections, digital libraries, computation | Ajax, multimedia, mobile databases |
| **ET3-tf** | Application, dimensional modeling, data warehouses | Generation, xml data, query aware | Algorithms, object reporitories, large scale databases | Web collections, computations, clustering | Ajax, multimedia, mobile databases |
| **ET1-mdeg** | Application, dimensional modeling, data warehouses | Generation, xml data, query aware | Algorithms, scale, ontologies | Web collections, digital libraries, clustering | Ajax, mdbc, mobile databases |
| **ET1-medeg** | Application, dimensional modeling, data warehouses | Generation, xml data, query aware | Algorithms, object reporitories, ontologies | Web collections, digital libraries, clustering | Ajax, mdbc, mobile databases |
| **ET1-deg/tf** | Drill-across queries, data warehousing, data oriented | Optimizing continuous queries, xml documents, sensor networks | Large databases, efficient approach, framework | Computing page reputation, author name disambiguation, hypergraph model | Adaptive join algorithm, semantic-based predicates implication, extreme restrictions |
| **ET2-deg/tf** | Drill-across queries, data warehousing, multidimensional modeling | Optimizing continuous queries, xml documents, adaptive aggregation algorithm | Large databases, efficient approach, based decisions | Computing page reputation, hypergraph model, heuristc-based hierarchical clustering method | Adaptive join algorithm, semantic-based predicates implication, towards efficient horizontal multimedia database fragmentation |
| **ET3-deg/tf** | Drill-across queries, data warehousing, multidimensional modeling | Optimizing continuous queries, xml documents, adaptive aggregation algorithm | Large databases, efficient approach, based decisions | Computing page reputation, hypergraph model, heuristc-based hierarchical clustering method | Adaptive join algorithm, semantic-based predicates implication, towards efficient horizontal multimedia database fragmentation |

| Ano | 2007 | | | | 2008 |
|---|---|---|---|---|---|
| **Tópico** | Query, clustering, algorithm, similarity, metric | Geographical, similarity, analogy, warehousesm, application | Databases, cml, detection, native, schema | Temporal, pattern, mining, hybrid, reduced | Queries, pagination, hash, results, indexing |
| **Rótulos** | | | | | |
| **ET1-tf** | Similarity queries, metric spaces, clustering | Analogy, similarity, application domain | Xml, detection, databases | Pattern mining, temporal data, opinion mining | Queries, results, dbms |
| **ET2-tf** | Similarity queries, metric spaces, algorithm | Analogy, similarity, geomdql | Xml, databases, schema | Pattern mining, temporal data, time domain | Queries, indexing, dbms |
| **ET3-tf** | Similarity queries, clustering, algorithm | Analogy, geomdql, geographical data warehouses | Xml schema, detection, databases | Pattern mining, temporal data, time domain | Queries, results, dbms |
| **ET1-mdeg** | Similarity queries, metric spaces, clustering | Analogy, similarity, geographical data warehouses | Xml schema, xml, databases | Pattern mining, temporal data, opinion mining | Queries, results, comprehensiveness |
| **ET1-medeg** | Similarity queries, metric spaces, clustering | Analogy, similarity, geographical data warehouses | Xml schema, Xml, databases | Pattern mining, temporal data, opinion mining | Queries, results, comprehensiveness |
| **ET1-deg/tf** | Grid-based clustering algorithm, evolucionary density, metric spaces | Querying geographical data warehouses, new approach, similarity queries | Xml schema evolution, native xml databases, preserves validity | Mining temporal relational patterns, reduced star-cubing approach, domains | Minimal perfect hash functions, indexing internal memory, queries |
| **ET2-deg/tf** | Grid-based clustering algorithm, evolucionary density, constrained aggregate similarity queries | Querying geographical data warehouses, new approach, neighborhood graphs | Xml schema evolution, native xml databases, embedding similarity joins | Mining temporal relational patterns, reduced star-cubing approach, mdag-cubing | Minimal perfect hash functions, indexing internal memory, results |
| **ET3-deg/tf** | Grid-based clustering algorithm, evolucionary density, constrained aggregate similarity queries | Querying geographical data warehouses, new approach, neighborhood graphs | Xml schema evolution, native xml databases, embedding similarity joins | Mining temporal relational patterns, reduced star-cubing approach, mdag-cubing | Minimal perfect hash functions, indexing internal memory, results |

| Ano | 2008 | | | | 2009 |
|---|---|---|---|---|---|
| **Tópico** | Cube, xml, computing, dimension, model | Record, object digital, data, image | Integration. Workflow, ontology, control, approach | Software, mining, application, reporitories, impact | Author, network, disambiguation, method, deduplication |
| **Rótulos** | | | | | |
| **ET1-tf** | Model, xml, data cube | Data, estimation, automatic | Bioinformatics, distribution, data | Application, mining, case | Networks, deduplication, disambiguation |
| **ET2-tf** | Model, computing approach, data cube | Data, image retrieval, digital objects | Bioinformatics, workflows, ontology integration | Application, case, software | Networks, method, deduplication |
| **ET3-tf** | Model, data cube, xml | Data, image retrieval, digital objects | Bioinformatics, data workflows, ontology integration | Application, software, minig | Networks, deduplication, evaluating |
| **ET1-mdeg** | Model, xml, data cube | Data, estimation, digital objects | Bioinformatics, distribution, data | Application, mining, case | Networks, deduplication, disambiguation |
| **ET1-medeg** | Model, xml, data cube | Data, image retrieval, digital objects | Bioinformatics, data workflows, ontology integration | Application, mining, case | Networks, deduplication, disambiguation |
| **ET1-deg/tf** | Computing data cubes, sequential mcq approach, xml instance level integration | Genetic programming approach, record deduplication, impact | Emerging ontologies, integration, approach | Mining software repositories, impact analysis, agroindustry | Author name disambiguation, digital libraries, automatic selection |
| **ET2-deg/tf** | Computing data cubes, sequential mcq approach, single graph paths | Genetic programming approach, record deduplication, parameters setup | Emerging ontologies, integration, bioinformatics | Mining software repositories, impact analysis, connectionblock algorithm | Author name disambiguation, digital libraries, genetic programming |
| **ET3-deg/tf** | Computing data cubes, sequential mcq approach, single graph paths | Genetic programming approach, record deduplication, parameters setup | Emerging ontologies, integration, bioinformatics | Mining software repositories, impact analysis, connectionblock algorithm | Author name disambiguation, digital libraries, genetic programming |

| Ano | 2009 | | | 2010 | |
|---|---|---|---|---|---|
| **Tópico** | Web, images, engine, search, documents | Mining, series, medical, solap, modeling | Compute, processing, bases, environment, clouds | Query, performance, rule, spatial, function | Rank, learning, discriminative, features, clickthrough |
| **Rótulos** | | | | | |
| **ET1-tf** | Recovery, images, web | Mining climate, remote sensing, time series | Consultations, processing, bases | Query, data storage, spatial data | Wcl2r, l2r, features |
| **ET2-tf** | Recovery, web, documents | Mining climate, discover, remote sensing | Consultations, bases, clouds | Query level, data storage, spatial data | Wcl2r, l2r, clickthrough data |
| **ET3-tf** | Recovery, documents, images | Mining climate, time series, discover | Consultations, environment, processing | Query level, data storage, spatial data | Wcl2r, clickthrough data, representative learning |
| **ET1-mdeg** | Recovery, images, web | Mining climate, remote sensing, time series | Consultations, processing, bases | Query, data storage, spatial data | Wcl3r, l2r, clickthrough data |
| **ET1-medeg** | Recovery, images, web | Mining climate, remote sensing, time series | Consultations, processing, bases | Query, query level, data storage | Wcl2r, clickthrough data, representative learning |
| **ET1-deg/tf** | Automatic classification, robust documents temporarily, genetic programming | Relevant climate patterns, similarity searching, incorporating metric access methods | Geostatistical data using partial replication, high performance, medical records | uses query-level rules, spatial data warehouse schemas, multiple query-level functions | clickthrough data, benchmark collections, rank L2R algorithms |
| **ET2-deg/tf** | Automatic classification, robust documents temporarily, web using multiple textual evidence | Relevant climate patterns, similarity searching, oracle database | Geostatistical data using partial replication, high performance, multifaceted analysis | uses query-level rules, spatial data warehouse schemas, data storage | clickthrough data, benchmark collections, rank L2R algorithms |
| **ET3-deg/tf** | Automatic classification, robust documents temporarily, web using multiple textual evidence | Mining climate, remote sensing, time series | Consultations, processing, bases | Data storage, uses query-level rules, spatial data warehouse schemas | clickthrough data, benchmark collections, rank L2R algorithms |

| Ano | 2010 | | | | 2011 |
|---|---|---|---|---|---|
| **Tópico** | Xml, search, temporal, keyword, phrasal | Source, local, ontology, field, user | Web, online, databases, collection, social | Classifier, mining, vídeo, classification, networks | Collection, image, library, digital, large |
| **Rótulos** | | | | | |
| **ET1-tf** | Xml, query processing, keyword search | Domain ontology, application ontology, local ontology | Content, online databases, web | Users, classifier, general | Image databases, digital library, image collections |
| **ET2-tf** | Query processing, keyword search, search engine | Domain ontology, application ontology, local ontology | Web forms, online databases, content | Users, protein classification, video spammers | Image databases, digital library, image collections |
| **ET3-tf** | Query processing, keyword search, search engine | Domain ontology, application ontology, local ontology | web forms, Content, online databases, | Users, protein classification, video spammers | Image databases, digital library, image collections |
| **ET1-mdeg** | Xml, query processing, keyword search | Domain ontology, application ontology, local ontology | Content, web, online databases | Users, classifier, protein classification | Image databases, digital library, image collections |
| **ET1-medeg** | Xml search engine, keyword search, query processing | Domain ontology, application ontology, local ontology | Web forms, online databases, test collection | Users, classifier, protein classification | Image databases, digital library, image collections |
| **ET1-deg/tf** | XML keyword search engines, intra-query parallel processing, identifying temporal constraints | user's preference hierarchy, generate application ontologies, corresponding local ontology | test collection, diferent storage configurations, support eficient query processing | general purpose classifier, protein classification problem, general classification method | large image collections, heterogeneous image databases, content-based image retrieval |
| **ET2-deg/tf** | XML keyword search engines, intra-query parallel processing, identifying temporal constraints | user's preference hierarchy, generate application ontologies, corresponding local ontology | Test collection, legitimate users, diferent storage configurations | general purpose classifier, protein classification problem, general classification method | heterogeneous image databases, image manipulation software, small image collections whereas |
| **ET3-deg/tf** | XML keyword search engines, intra-query parallel processing, identifying temporal constraints | user's preference hierarchy, generate application ontologies, corresponding local ontology | Test collection, legitimate users, diferent storage configurations | general purpose classifier, protein classification problem, general classification method | content-based image retrieval, image manipulation software, small image collections whereas |

| Ano | 2011 | | | | |
|---|---|---|---|---|---|
| **Tópico** | Document, research, classification, challenge, effect | Attribute, selection, improve, keyword, approach | Cube, workflow, olap, engine, data | Ontology, reuse, temporal, tool, ufo | Neighbor, context, knn, nearest, operator |
| **Rótulos** | | | | | |
| **ET1-tf** | Mining algorithm, document classification, temporal | Learning, atribute, keyword | Olap, workflow, data | Ontologies, domain, knowledge | Knn, query execution, k-nearest |
| **ET2-tf** | Mining algorithm, document classification, temporal evolution | Keyword, Learning approach, atribute selection | Olap, workflow execution, data intensive | Ontologies, knowledge, domain ontology | k-nearest, knn, comparison operator |
| **ET3-tf** | Mining algorithm, document classification, temporal evolution | Keyword, machine learning approach, atribute selection | Workfloe execution, olap, data intensive | Ontologies, knowledge, domain ontology | Knn, k-nearest, comparison operator |
| **ET1-mdeg** | Mining algorithm, document classification, temporal evolution | Keyword, machine learning, classification task | Olap, workfloe execution, data intensive | Ontologies, knowledge, domain | Knn. K-nearest, comparison operator |
| **ET1-medeg** | Mining algorithm, document classification, temporal evolution | Machine learning, attribute selection, select keyword | Olap, workflow execution, data intensive | Ontologies, knowledge, domain ontology | Knn. K-nearest, comparison operator |
| **ET1-deg/tf** | temporal evolution, important research topic, large classification problems | new approach relies, lazy learning approach, lazy attribute selection technique | data processing tasks, local database engine, data-intensive workflows | ontology tools, specific domain, domain ontology | attribute comparison operators, k-nearest neighbor operators, k-nearest neighbor query |
| **ET2-deg/tf** | important research topic, large classification problems, graph mining algorithm | lazy attribute selection technique, new attribute selection strategy, machine learning strategies | data processing tasks , ant colony optimization, data-intensive workflows | ontology tools, specific domain, domain ontology | regular k-nn operation, k-nn queries attribute comparison operators, k-nearest neighbor operators, k-nearest neighbor query |
| **ET3-deg/tf** | Graph mining algorithm, important research | lazy attribute selection technique, new attribute | data processing tasks , ant colony optimization, data- | software process ontology, ontology tools, hinder | attribute comparison operators, k-nearest neighbor operators |

| | | | | | |
|---|---|---|---|---|---|
| | topic, large classification problems | selection strategy, machine learning strategies | intensive workflows | ontology comprehension | |

| **Ano** | **2012** | | | | |
|---|---|---|---|---|---|
| **Tópico** | Graph, compute, kernel, accuracy, better | Cluster, level, metric, internal, k-nearest | File, computer, memory, performance, local | Consumption, repository, location, resource, using | Prediction, author, name, record, correct |
| **Rótulos** | | | | | |
| **ET1-tf** | Graph data, quality metrics, internal density | Cluster vertex, cluster, large | File system, such memory, device data | Etl, index structures, data warehouse | Author name, machine learning, training |
| **ET2-tf** | Graph data, quality metrics, internal density | Cluster, large datasets, classification step | File system, flash memory, such memory | Etl, index structures, data warehouse | Author name, machine learning, user |
| **ET3-tf** | Graph data, quality metrics, internal density | Cluster vertex, large datasets, classification step | File system, flash emmory, volatile memory área | Extensible framework, index structures, data warehouse | Author name, machine learning, user relevance |
| **ET1-mdeg** | Graph, quality metrics, new method | Cluster, large, datasets | File system, device data, flash memory | Etl, main objectives, data warehouse | Author name, user relevance feedback, machine learning |
| **ET1-medeg** | Graph data, real graph databases, internal density | Cluster, large datasets, machine learning | File system, device data, efficient file systems | Etl, data warehouses, index structures | Machine learning, author name, user relevance feedback |
| **ET1-deg/tf** | traditional quality metrics, powerful data management algorithms, graph data management | uci machine learning, various performance improvements, large scale datasets | Persistent memory area, efficient file systems, flash file system | query execution history, data integration processes, data warehouses | machine learning techniques, exists citation records, overall disambiguation effectiveness improves |
| **ET2-deg/tf** | traditional quality metrics, powerful data management algorithms, graph data management | uci machine learning, high computational cost, large scale datasets | Persistent memory area, efficient file systems, herdware platform | Data warehouses, query execution, extensible framework | name ambiguity, author names, experimental evaluation |
| **ET3-deg/tf** | traditional quality metrics, good external sparsity evaluation metric, graph data management | uci machine learning, high computational cost, large scale datasets | Persistent memory area, efficient file systems, flash file system | query execution history, data integration processes, data warehouses | overall disambiguation effectiveness improves, machine learning techniques, user relevance feedback |

| Ano | 2012 | | |
|---|---|---|---|
| **Tópico** | Model, crime, hub, activity, complex | Social, web, analysis, human, networks | Query, similar, tool, context, process |
| **Rótulos** | | | |
| **ET1-tf** | Collaboration, malicious activities, reports | Data analysis, social netowrks, small data | Query returns, similarity, previous |
| **ET2-tf** | Collaboration, malicious activities, complex networks | Data analysis, social netowrks, small data | Query return, search algorithm, neighbor query |
| **ET3-tf** | Collaboration, malicious activities, complex networks | Data analysis, social netowrks, small data | Query return, search algorithm, neighbor query |
| **ET1-mdeg** | Collaboration, reports, hubs | Data, social networks, specific sports | Query, similarity, knn |
| **ET1-medeg** | Complex networks, malicious activities, hub users | Social networks, data analysis, temporal factors | Query return, knn, Neighbor query |
| **ET1-deg/tf** | Complex networks, such non-hub users, deviations arising from malicious activity | basic human activities, data analysis, social networks concern different types | processing k-nearest neighbors, knn queries, algorithm |
| **ET2-deg/tf** | such non-hub users, deviations arising from malicious activity, bipartite network model | Data analysis, ignores temporal factors, basic human activities | processing k-nearest neighbors, knn queries, search algorithm |
| **ET3-deg/tf** | such non-hub users, deviations arising from malicious activity, bipartite network model | basic human activities, understanding information diffusion, Sport social networks | processing k-nearest neighbors, knn queries, search algorithm |