Evolução dos Temas de Interesse do SBBD ao Longo dos Anos

Anderson Kauer, Viviane Moreira

Instituto de Informática – UFRGS, Brazil
{aukauer, viviane}@inf.ufrgs.br

Abstract. A mineração temporal de textos (temporal text mining - TTM) é uma técnica 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. Além disso, a mineração temporal de textos é útil em ferramentas de sumarização e descoberta de tendências. Neste trabalho, aplicamos TTM para analisar a evolução dos temas abordados pelos artigos publicados no Simpósio Brasileiro de Bancos de Dados (SBBD) ao longo dos anos. Para tal, coletamos os abstracts dos artigos publicados entre 1989 e 2012. Os resultados mostram tendências interessantes na distribuição dos temas ao longo do tempo.

Categories and Subject Descriptors: H.3.3 [**Information Storage and Retrieval**]: Clustering

Keywords: clustering, evolução temporal, mineração de textos

1. INTRODUÇÃO

Em diferentes domínios de aplicações, existe a necessidade de se conhecer os principais temas discutidos em coleções de texto. Um cenário bastante comum consiste em bases de artigos científicos, onde em determinado ponto no tempo os temas de interesse são modificados. Isto faz com que alguns assuntos tornem-se mais expressivos, sejam descontinuados, ou até mesmo se especializem em novos temas. A mineração temporal de textos (*temporal text mining* – TTM) é uma técnica que consiste em identificar estas estruturas latentes como responsáveis por definir o assunto ou modelo gerador dos documentos [Mei and Zhai 2005], assim é possível identificar como o estudo de um tópico influenciou o estudo de um outro tópico em épocas diferentes.

Neste trabalho, aplicamos o método para TTM que foi introduzido por [Mei and Zhai 2005] como forma de identificar automaticamente a evolução dos temas abordados no Simpósio Brasileiro de Bancos de Dados (SBBD) ao longo dos anos. O método utilizado consiste na criação de um modelo probabilístico que visa descobrir os temas latentes nas coleções em cada edição e identificar as relações entre estes temas ao longo dos anos.

O restante deste trabalho está organizado da seguinte maneira: na Seção 2 apresentamos uma breve descrição dos trabalhos que compõem o estado da arte. Na Seção 3 descrevemos os detalhes do método implementado e na Seção 4 são avaliados os resultados obtidos nos experimentos. Finalmente, na seção 5 são discutidos os resultados e apontamos algumas direções para trabalhos futuros.

2. TRABALHOS RELACIONADOS

Muitos aspectos da TTM têm sido utilizados em outras áreas de pesquisa. Inicialmente, a TTM estava fortemente relacionada à sumarização temporal [Allan et al. 2001]. Este artigo, embora relacionado com sumarização, tem como diferencial a identificação das características implícitas no texto.

A detecção e o rastreamento de temas (*topic detection and tracking*–TDT) [Kaur and Gupta 2012] concentra-se em descobrir e extrair informações comuns a partir de um conjunto de documentos. A relação com este trabalho está na identificação da estrutura latente da coleção, entretanto a TDT não considera os padrões temporais e evolucionários da coleção.

Sipos et. al. [Sipos et al. 2012] desenvolveram um trabalho bastante semelhante ao nosso propósito, entretanto, o objetivo é identificar a influência de documentos e autores com o passar dos anos, enquanto que não são consideradas as transições entre temas implícitos na coleção.

Por fim, algoritmos de agrupamento para coleções de texto têm ganhado bastante atenção nos últimos anos devido à capacidade de identificar estruturas latentes dos documentos [Zhai et al. 2004]. Esta abordagem é interessante por identificar características comuns e raras em diferentes coleções. Adicionalmente é possível identificar características comuns entre os temas conforme descrito em [Mei and Zhai 2005] como forma de identificar as transições e intensidade em cada período de tempo.

3. TTM SOBRE O SBBD

Para o desenvolvimento deste trabalho, seguiu-se a metodologia abordada em [Mei and Zhai 2005]. Onde, por definição, existe conjunto de coleções $\{C_1, C_2, \ldots, C_n\}$, onde cada coleção $C_i = \{d_1, d_2, \ldots, d_n\}$ e cada documento d_j é representado por uma sequência de palavras que compõem o vocabulário $V = \{w_1, w_2, \ldots, w_{|V|}\}$, onde cada w_i é uma palavra.

Neste contexto, um *tema* θ em uma coleção C_i segue uma distribuição probabilística de palavras que caracterizam um tópico. Esta distribuição permite que palavras que compartilham um determinado tema estejam próximas.

Assim, cada coleção C_i é composta por um conjunto de temas $\Theta_i = \{\theta_1, \theta_2, \ldots, \theta_n\}$, que também podem ser considerados os modelos responsáveis por gerar os documentos. Cada documento d_j tem probabilidade $\pi_{j,k}$ de ter sido gerado pelo tema θ_k e portanto $\sum_{k=0}^{n} \pi_{j,k} = 1$.

3.1 Extração dos temas

A extração dos temas em cada coleção C_i seguiu um modelo probabilístico de mistura que foi descrito em [Zhai et al. 2004] e, posteriormente, adaptado para TTM [Mei and Zhai 2005]. Este método calcula a distribuição das palavras em todas as coleções (Eq. 1), gerando o modelo conhecido como *background text* (θ_B), para então estimar um conjunto de modelos responsáveis pela distribuição das palavras em cada documento.

$$\hat{p}(w|\theta_B) = \frac{\sum_{i=1}^{m} \sum_{d \in C_i} c(w,d)}{\sum_{i=1}^{m} \sum_{d \in C_i} \sum_{w' \in V} c(w',d)} \tag{1}$$

onde $c(w,d)$ indica o número de ocorrências da palavra w no documento d, m é o número de coleções e w' é uma palavra pertencente ao vocabulário V.

O ajuste dos parâmetros pode ser feito utilizando uma variante do algoritmo *Expectation Maximization* (EM) [Dempster et al. 1977]. A utilização de λ_B como uma constante tem como objetivo tornar as palavras mais discriminativas em relação à coleção, reduzindo assim a influência de palavras pouco informativas.

Seguindo a definição do algoritmo EM, cada iteração ocorre em dois passos: no passo E (*expectation*) são estimadas as variáveis ocultas $\{z_{d,w}\}$ e suas probabilidades $p(z_{d,w} = j)$ indicando que a palavra w no documento d foi gerada a partir do tema j:

$$p(z_{d,w} = j) = \frac{\pi_{d,j}^{(n)} p^{(n)}(w|\theta_j)}{\sum_{j'=1}^{k} \pi_{d,j'}^{(n)} p^{(n)}(w|\theta_{j'})} \tag{2}$$

$$p(z_{d,w} = B) = \frac{\lambda_B p(w|\theta_B)}{\lambda_B p(w|\theta_B) + (1 - \lambda_B) \sum_{j=1}^{k} \pi_{d,j}^{(n)} p^{(n)}(w|\theta_j)} \tag{3}$$

onde $p(z_{d,w} = B)$ indica a probabilidade da palavra w seguir a distribuição das palavras na coleção e n é o número da iteração.

Em seguida, no passo M (*maximization*) os parâmetros são atualizados:

$$\pi_{d,j}^{(n+1)} = \frac{\sum_{w \in V} c(w,d) p(z_{d,w} = j)}{\sum_{j'=1}^{k} \sum_{w \in V} c(w,d) p(z_{d,w} = j')} \tag{4}$$

$$p^{(n+1)}(w|\theta_j) = \frac{\sum_{i=1}^{m} \sum_{d \in C_i} c(w,d)(1 - p(z_{d,w} = B)) p(z_{d,w} = j)}{\sum_{w' \in V} \sum_{i=1}^{m} \sum_{d \in C_i} c(w',d)(1 - p(z_{d,w'} = B)) p(z_{d,w'} = j)} \tag{5}$$

onde $p(w|\theta_j)$ indica a probabilidade da palavra w ser gerada a partir do tema j.

Uma característica deste método é a garantia de convergência que maximiza a verossimilhança dos dados. Assim, o conjunto de parâmetros $\Lambda = \{\theta_j, \pi_{d,j} | d \in C_i, 1 \leq j \leq k\}$ pode ser avaliado ao final de cada iteração como:

$$\mathcal{L} = \sum_{d \in C_i} \sum_{w \in V} \left[c(w,d) \log(\lambda_B p(w|\theta_B) + (1 - \lambda_B) \sum_{j=1}^{k} (\pi_{d,j} p(w|\theta_j))) \right] \tag{6}$$

A utilização deste método exige que seja definido o número de temas em cada coleção C_i. Para isto, foram analisadas, de maneira empírica, as probabilidades $\pi_{d,j}$ em cada coleção de forma que cada um dos documentos pertencesse a um determinado tema com uma alta probabilidade (i.e., maior do que 0.7). Também é necessária a definição da constante λ_B, que na prática serve como o fator de discriminação entre as palavras. Neste contexto, λ_B controla o ruído existente na coleção. [Zhai et al. 2004] sugerem valores de λ_B entre 0.9 e 0.95 . Em nossos experimentos utilizou-se $\lambda_B = 0.95$ juntamente com remoção de palavras comuns (*stop-words*) e *Light Stemming* que consiste em remover os sufixos indicativos de plural, gerúndio e passado (*s, ing, ed*). Também não foram utilizadas *keywords*, pois este trabalho visa identificar as características implícitas nos documentos.

3.2 Análise das Transições Evolucionárias

A análise de transições consiste em identificar a proximidade entre os temas de coleções diferentes. No caso do SBBD, seria a proximidade entre temas de edições diferentes. Considerando que cada tema em uma coleção pode ser representado por $\gamma = \langle \theta_j, C_i \rangle$, uma transição evolucionária ocorre quando a distância entre γ_1 e γ_2 for menor que um determinado limiar. A medida utilizada segue a distância de Kullback-Leibler [Cover and Thomas 1991] ajustada ao modelo de mistura [Mei and Zhai 2005] apresentado anteriormente. Assim a proximidade entre dois temas $D(\theta_2 \| \theta_1)$ é calculada como segue.

$$D(\theta_2 \| \theta_1) = \sum_{i=1}^{|V|} p(w_i|\theta_2) \log \frac{p(w_i|\theta_2)}{p(w_i|\theta_1)} \tag{7}$$

onde V representa a união das palavras pertencentes a θ_1 e θ_2. Temas diferentes compartilham poucas palavras, assim utilizou-se a metodologia proposta em [Pinto et al. 2007] para evitar que temas com poucas palavras em comum obtivessem uma distância muito baixa.

Finalmente, são consideradas como transições evolucionárias todos os pares de temas que atingiram distância inferior ao limiar ξ. Assim, a definição deste limiar depende de análise de cada uma das combinações possíveis. Em nossos experimentos definiu-se empiricamente a utilização de $\xi = 60$.

4. EXPERIMENTOS E RESULTADOS

Inicialmente foram coletados os *abstracts* de artigos completos de todas as edições do SBBD (de 1986 a 2012). Assim, foram obtidas as coleções C_i necessárias. Um dos problemas encontrados nesta etapa foi a heterogeneidade entre os documentos: muitas edições antigas estão disponíveis somente na versão impressa; alguns artigos possuem resumo somente em português ou espanhol. Os artigos impressos foram digitalizados manualmente. Como focamos a análise nos *abstracts* em inglês, as edições de 1986 a 1988 foram descartadas. Portanto, a análise foi realizada sobre 475 *abstracts* publicados entre 1989 e 2012. Este conjunto de documentos foi particionado em 23 intervalos de tempo, onde cada intervalo representa uma edição do SBBD. Cada edição é composta por um número diferente de palavras e de documentos, este fato implica na utilização de um número variável de temas para cada uma destas coleções.

Para definir o número de temas foi avaliada a probabilidade $\pi_{d,j}$. Valores muito baixos para todos os documentos em um determinado tema implicam que um número maior de temas é necessário. Outro aspecto analisado foi o conjunto de temas final: visto que temas muito semelhantes indicam que existe um excesso de temas. A Tabela I apresenta as cinco primeiras palavras com maior probabilidade em cada um dos temas identificados. Devido ao grande número de coleções, são apresentados somente quatro temas em cada década.

Inicialmente, nota-se que cada um dos temas trata de algum aspecto específico da área de banco de dados. Em 1989 os principais temas eram modelagem em Redes de Petri, e bases de conhecimento. Em 1990 são identificados mais temas relacionados operações sobre visões (*views*), estruturas de janelas de transações, representação de regras em banco de dados, expressão de consultas e armazenamento de imagens. Em 1991 são apresentados trabalhos relacionados ao modelo ER Estendido, hierarquias tolerantes a falhas, banco de dados temporal e técnicas declarativas para otimização de consultas. Em 1992 são abordados modelagem e design de dados, expressão e otimização de consultas, e representação de objetos complexos.

Nove anos depois, em 1999 são apresentados trabalhos sobre banco de dados distribuídos, modelos multi-dimensionais, técnicas para criação de índices e persistência. Em 2000 os principais tópicos estão relacionados a dados temporais, transferência de dados, *data warehouse*, processamento e recuperação de e-mails, processamento paralelo de consultas e restrições de integridade. Em 2001 os trabalhos abordaram técnicas para criação de índices, dados temporais, *data mining*, integração de dados e versionamento de esquemas. Em 2002 os temas foram criptografia de dados, performance de consultas, dados semi-estruturados, banco de dados orientado a objetos, recuperação de informação multilíngue e ferramentas de auxílio à tomada de decisão.

Nos anos mais recentes, os temas mais relevantes em 2009 estão relacionados a consultas utilizando inferência probabilística, motores de busca e base de dados centralizadas, revisões no estado da arte, aprendizagem de máquina, mineração de dados e computação em nuvem. Em 2010, os temas mais relevantes estão relacionados à ordenação e recuperação de consultas, *clickthrough information*, identificação de *phrasal terms*, interpretação de palavras-chave, web semântica, classificação e recuperação de vídeo. Em 2011 o trabalhos abordam recuperação e processamento de imagens, algoritmos de classificação e ordenação, *data-intensive workflows* e ontologias. Finalmente, em 2012, as principais contribuições estão relacionadas a banco de dados orientado a grafos, algoritmos de *clustering*, *hardware*, eficiência de consumo de energia, identificação de referências em artigos e identificação de usuários maliciosos.

A partir dos temas extraídos, utilizou-se a distância de KL, conforme descrito na seção 3.2, para identificar as transições evolucionárias. A Figura 1 apresenta o grafo com as principais transições mostrando o fluxo entre os temas ao abordados ao longos dos anos. Praticamente não há transições significativas até 1996. Os temas tratados nas diferentes edições do SBBD destes anos são mais dissimilares entre si. A maior parte das transições aparece a partir de 1996. O Tema 5 (dados

Tabela I. Temas extraídos dos abstracts do SBBD

	Tema 1	Tema 2	Tema 3	Tema 4	Tema 5	Tema 6
1989	modell 0,155 nets 0,124 petri 0,124 entit 0,112 informalit 0,062	eiti 0,075 base 0,062 oms 0,050 prolog 0,049 datalog 0,049	knowledg 0,100 coupl 0,048 kee 0,034 dbs 0,034 krisy 0,033			
1990	view 0,240 updat 0,102 operation 0,054 constraint 0,048 directl 0,038	window 0,159 featur 0,057 pixel 0,046 pyrami 0,046 insid 0,044	rule 0,091 languag 0,087 rdl 0,054 program 0,075 variabl 0,050	imag 0,106 eds 0,038 cod 0,033 manager 0,032 express 0,027		
1991	eer 0,117 typ 0,096 lead 0,066 entit 0,065 extend 0,047	hierarch 0,096 answer 0,091 cooperativ 0,048 cobas 0,044 fault 0,044	temporal 0,113 stat 0,076 condition 0,056 disk 0,042 time 0,042	optimizer 0,085 dbms 0,048 commit 0,035 cope 0,033 purpos 0,021		
1992	application 0,155 aristot 0,055 composit 0,053 target 0,051 designer 0,045	catalog 0,106 standar 0,088 design 0,082 assessment 0,053 purpos 0,037	imag 0,086 class 0,070 complexit 0,066 languag 0,055 circumvent 0,036	step 0,058 appropriat 0,041 attach 0,035 workbench 0,033 cooperation 0,031	object 0,069 iql 0,052 dbm 0,034 interactiv 0,027 formulation 0,026	lock 0,044 expert 0,027 relat 0,026 tabl 0,026 transient 0,024
1999	migration 0,086 execution 0,068 network 0,060 ral 0,052 dimension 0,052	mdbs 0,141 transaction 0,098 local 0,088 multidatabas 0,087 failur 0,068	activ 0,065 data-driven 0,063 essentiall 0,062 rul 0,049 event-action 0,042	modul 0,074 snep 0,069 dimensional 0,045 dimension 0,042 dvr 0,042	index 0,059 exampl 0,034 dedy 0,033 hierarch 0,032 conventional 0,030	persistenc 0,044 action 0,035 reflection 0,027 uncertaint 0,026 fuzz 0,020
2000	temporal 0,155 kess 0,046 program 0,038 metadata 0,031 biodiversit 0,029	client 0,103 broadcast 0,081 mobil 0,064 protocol 0,051 listen 0,032	acces 0,077 spatial 0,057 warehous 0,054 factor 0,037 internet 0,032	mbe 0,071 visual 0,065 email 0,051 interfac 0,038 not 0,035	polylin 0,057 pair 0,043 approximation 0,035 parallel 0,029 intersect 0,028	attribut 0,046 constraint 0,045 corporation 0,035 formula 0,032 dimension 0,025
2001	index 0,108 tabl 0,107 join 0,102 vector 0,094 bitmap 0,051	temporal 0,134 join 0,129 r-tre 0,085 deby 0,044 semistructur 0,044	imag 0,145 itemset 0,082 pelican 0,052 maximal 0,050 frequent 0,043	mediator 0,094 juridical 0,043 repartition 0,043 generation 0,041 thesauru 0,029	medical 0,040 evolution 0,040 gis 0,038 collection 0,024 diseas 0,023	nam 0,041 attribut 0,035 olap 0,029 word 0,022 file 0,020
2002	protocol 0,086 extraction 0,078 encrypt 0,062 decipher 0,047 ontolog 0,046	mart 0,062 sourc 0,061 optimization 0,049 dig 0,045 pre-model 0,045	mediat 0,079 accurac 0,056 reliabl 0,050 xmls 0,047 correspondenc 0,042	odbm 0,063 dsmio 0,063 prism 0,050 revers 0,036 association 0,028	parallel 0,034 join 0,034 cross-languag 0,031 evaluat 0,028 odmg 0,027	mdbc 0,049 decision 0,039 prioritization 0,031 vsm 0,029 host 0,029
2009	probabilistic 0,196 inferenc 0,079 sql 0,048 safe 0,048 offlin 0,032	engin 0,090 cost 0,089 olap 0,055 pargr 0,052 centraliz 0,048	solution 0,078 challeng 0,048 section 0,040 art 0,039 dimensional 0,039	record 0,082 block 0,060 exampl 0,053 oracl 0,050 genetic 0,043	survival 0,040 solap 0,039 climat 0,032 xbrl 0,031 smtm 0,024	tutorial 0,045 subsumption 0,029 cover 0,023 googl 0,021 functional 0,019
2010	rank 0,101 relevanc 0,091 function 0,085 query-level 0,075 query-sensitiv 0,075	wclr 0,133 collection 0,114 clickthrough 0,106 featur 0,070 discriminativ 0,053	phrasal 0,204 term 0,164 ontolog 0,069 stream 0,057 detect 0,047	search 0,087 ontolog 0,071 form 0,059 keywor 0,043 keyword-bas 0,040	credibilit 0,064 classification 0,050 spatial 0,035 preferenc 0,035 lsh 0,029	video 0,079 cliqu 0,049 sentiment 0,048 youtub 0,028 cooperativ 0,028
2011	imag 0,179 descriptor 0,118 attribut 0,082 effectivenes 0,054 lazy 0,035	k-nearest 0,102 neighbor 0,102 operator 0,086 predicat 0,048 rewrit 0,041	rank 0,108 aggregation 0,062 learn 0,057 iets 0,048 keyword 0,043	workflow 0,109 grasp 0,071 data-intensiv 0,048 aco 0,030 pbs 0,030	ontolog 0,087 effect 0,045 classifier 0,036 temporally 0,029 adc 0,027	trajector 0,047 bias 0,044 ontolog 0,025 question 0,020 sla 0,019
2012	graph 0,360 fingerprint 0,107 kernel 0,089 comput 0,044 min-hash 0,035	cluster 0,175 densit 0,091 graph 0,090 internal 0,056 metric 0,045	memor 0,183 flash 0,076 devic 0,055 comput 0,054 file 0,040	etl 0,122 consumption 0,043 location 0,038 repositor 0,034 dysto 0,033	prediction 0,032 author 0,030 pim 0,030 sport 0,030 busines 0,022	crim 0,026 maliciou 0,026 pais 0,026 pplocator 0,026 trac 0,026

geográficos) especializa-se em três diferentes temas: Temas 4 (mapeamento de objetos espaciais) e 5 (recuperação de imagens), ambos de 1997, e Tema 6 (*hypermedia*) em 1998. Os temas 5 (indexação) e 6 (persistência) de 1999 estão próximos ao Tema 5 (integração de dados) de 2001. Observa-se a formação de um fluxo a partir dos Temas 5 e 6 de 2001 para o Tema 5 (recuperação de informação multilíngue) em 2002, após, para o Tema 3 (*data mining*) em 2004 e posteriormente especializa-se nos Temas 2 (similaridade de objetos), 3 (mineração de dados temporais) e 5 (regras de associação e classificação) em 2005. Também há um fluxo do Tema 5 (dados temporais) em 2004 para o Tema 3 em 2005. Em 2006, os temas 4 (versionamento) e 5 (representação de objetos) deslocam-se para o Tema 5 (banco de dados geo-temporais) em 2007, mais tarde influenciando o Tema 5 (similaridade e recuperação de imagens) em 2008, e por fim o Tema 5 (duplicação e redundância de dados) em 2009. Em 2010, o Tema 4 (ontologias) especializa-se nos temas 3 (ordenação e seleção) e 5 (classificação e ontologias) em 2011. O Tema 5 (classificação) em 2010 passa pelo Tema 5 em 2011 e finaliza no Tema 4 (eficiência e alocação de recursos).

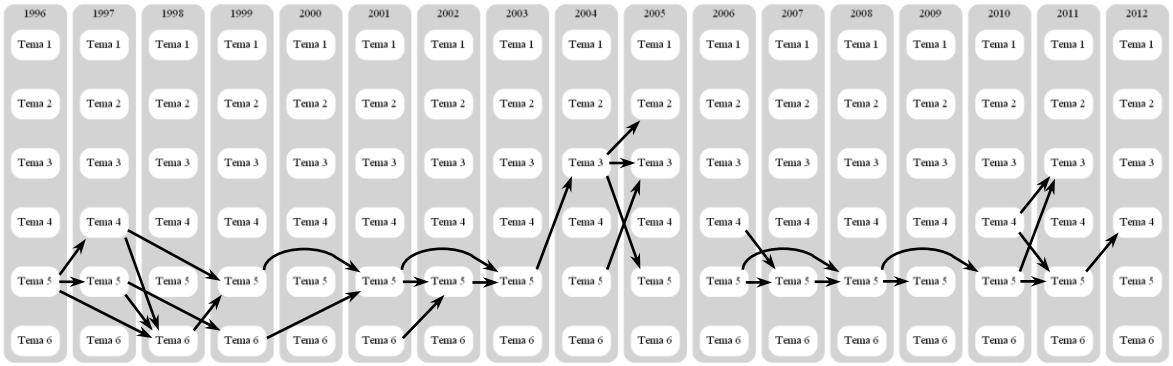

Figura 1. Transições entre os temas ao longo dos anos

5. CONCLUSÃO

A mineração temporal de dados é uma técnica que permite identificar as estruturas latentes em coleções de texto, facilitando o entendimento dos principais assuntos abordados em cada coleção. O método utilizado consiste formar agrupamentos (*clusters*) de palavras, identificando assim os temas da coleção. Neste trabalho, utilizamos os *abstracts* dos artigos publicados pelo SBBD para compor nossas coleções. Na etapa de classificação, identificou-se que os temas formados são bastante coerentes com o propósito do SBBD, isto permitiu uma análise das transições evolucionárias entre estes temas com o passar dos anos.

Os resultados mostram que, nos anos iniciais, os temas eram mais diferentes entre si a cada edição. Isto pode ser explicado pelo fato de que nos anos iniciais os artigos tratavam de tecnologias emergentes. Com o passar do tempo, essas tecnologias foram sendo combinadas gerando as transições evolucionárias.

Como trabalho futuro, pretendemos mensurar a intensidade dos temas em cada período. Assim será possível identificar os principais temas abordados. Seria interessante comparar os temas do SBBD com os de outro evento da área de banco de dados, como o VLDB, por exemplo.

Agradecimentos: O primeiro autor é aluno bolsista do CNPq. Este trabalho foi parcialmente financiado pelos projetos CNPq 478979/2012-6 e 480283/2010-9.

REFERÊNCIAS

Allan, J., Gupta, R., and Khandelwal, V. Temporal summaries of news topics. In *24th ACM SIGIR conference on Research and development in information retrieval*. ACM, New Orleans, Louisiana, USA, pp. 10–18, 2001.

Cover, T. M. and Thomas, J. A. *Elements of information theory*. Vol. 1. Wiley Online Library, New York, 1991.

Dempster, A. P., Laird, N. M., and Rubin, D. B. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)* vol. 39, pp. 1–38, 1977.

Kaur, K. and Gupta, V. A survey of topic tracking techniques. *International Journal of Advanced Research in Computer Science and Software Engineering* 2 (5): 384–393, 2012.

Mei, Q. and Zhai, C. Discovering evolutionary theme patterns from text: an exploration of temporal text mining. In *ACM SIGKDD international conference on Knowledge discovery in data mining*. KDD '05. ACM, Chicago, Illinois, USA, pp. 198–207, 2005.

Pinto, D., Benedí, J.-M., and Rosso, P. Clustering narrow-domain short texts by using the kullback-leibler distance. In *International Conference on Computational Linguistics and Intelligent Text Processing*. CICLing '07. Springer-Verlag, Mexico City, Mexico, pp. 611–622, 2007.

Sipos, R., Swaminathan, A., Shivaswamy, P., and Joachims, T. Temporal corpus summarization using submodular word coverage. In *ACM International Conference on Information and Knowledge Management*. CIKM '12. ACM, Maui, Hawaii, USA, pp. 754–763, 2012.

Zhai, C., Velivelli, A., and Yu, B. A cross-collection mixture model for comparative text mining. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '04. ACM, Seattle, WA, USA, pp. 743–748, 2004.