

Descoberta de Conhecimento em Textos através da Análise de Seqüências Temporais*

Stanley Loh^{1,2}, Leonardo Albernaz Amaral¹,
Leandro Krug Wives³, José Palazzo Moreira de Oliveira³

¹Escola de Informática - Universidade Católica de Pelotas (UCPEL)
Rua Félix da Cunha, 412 – CEP 96010-000 – Pelotas – RS – Brazil

²Depto. de Informática e Computação - Universidade Luterana do Brasil (ULBRA)
Av. Farroupilha, 8001 – CEP 92425-900 – Canoas – RS – Brazil

³Instituto de Informática – Universidade Federal do Rio Grande do Sul (UFRGS)
Caixa Postal 15.064 – CEP 91.501-970 – Porto Alegre – RS – Brazil

sloh@terra.com.br, sapo@vetorial.net, {wives,palazzo}@inf.ufrgs.br

***Abstract.** This paper presents a knowledge discovery process performed over texts, using a technique for analyzing time sequences. The technique was tested in medical records of patients admitted in a psychiatric hospital. Each time sequence is the set of records of one patient, chronologically sorted, representing evolution of the patient. Records are written by professionals using free text. One benefit of the process is to find symptoms and signals after a medicine is administrated to a patient.*

***Resumo.** Este artigo apresenta um processo de descoberta de conhecimento em textos, utilizando uma técnica de análise de seqüências temporais. Ela foi testada em prontuários médicos de pacientes internados em uma clínica psiquiátrica, onde cada seqüência é uma ordenação de registros de evoluções de um paciente e cada registro é escrito em textos livres por médicos e profissionais da área. Um dos benefícios consiste na identificação de certos sintomas e sinais após determinados remédios terem sido administrados.*

1. Introdução

Técnicas de descoberta de conhecimento são importantes para quem trabalha com um grande volume de informações, ajudando pessoas a descobrir conhecimento útil e novo, geralmente implícito, minimizando a sobrecarga de informações. Existem muitos trabalhos para descoberta de conhecimento em bancos de dados, isto é, sobre dados estruturados. Entretanto, a maioria das informações de uma organização estão codificadas em formato textual [TAN 99].

Para analisar textos, existem técnicas e ferramentas para Descoberta de Conhecimento em Textos [FEL 95] ou Text Mining [TAN 99]. Entretanto, a maioria dos trabalhos considera os textos como sendo atemporais, ou seja, formando um único

* Este trabalho é parcialmente apoiado por CNPq (projetos PERXML, Edital Universal nº 019/2004, proc. 475743/2004-0 e DIGITEX CTInfo 2005, Edital 11/2005, proc. 550.845/2005-4) e FAPERGS. Os prontuários médicos foram fornecidos pela Clínica Olivé Leite (Pelotas/RS, Brazil), onde são desenvolvidas pesquisas com suporte do FIDEPS (Fundo de Incentivo para o Desenvolvimento de Ensino e Pesquisa em Saúde – Ministério da Saúde).

conjunto sem tempo associado. Uma técnica temporal pode identificar padrões estatísticos em textos com tempo associado. Por exemplo, pode-se analisar uma seqüência de reclamações de clientes com o objetivo de verificar quais seqüência de ações as pessoas tomam durante este processo. Admitindo que os clientes registrem várias reclamações ao longo do tempo, cada seqüência é composta pelas reclamações de um único cliente e mantém a ordem cronológica em que foram registradas. Os resultados de uma tal análise permitiriam descobrir que os clientes geralmente iniciam reclamando de um produto ou serviço, tratam do atendimento, reclamam que não obtiveram resposta e terminam cancelando o serviço ou migrando para um concorrente.

Este artigo apresenta um processo de descoberta de conhecimento em textos utilizando uma técnica de análise de seqüências temporais em conceitos extraídos de textos. O processo recebe como entrada diversas seqüências, que são conjuntos de textos ordenados cronologicamente. O objetivo é encontrar correlações ou dependências entre características dos textos, de um texto para outro da mesma seqüência, mas avaliando a probabilidade condicional no conjunto de todas as seqüências. As características são conceitos extraídos dos textos por métodos de classificação que analisam suas palavras. A análise das seqüências temporais permite analisar a dependência entre conceitos ao longo do tempo, isto é, se um conceito condiciona a aparição de outro conceito no futuro.

A técnica foi testada em prontuários médicos de pacientes internados em uma clínica psiquiátrica, onde cada seqüência temporal é uma ordenação de registros de evoluções de um mesmo paciente, e cada registro é escrito em textos livres por médicos e profissionais da área. O artigo apresenta avaliações formais da técnica, mas também discute o conhecimento descoberto a partir deste processo, como por exemplo a aparição de certos sintomas após certos remédios terem sido administrados.

2. Trabalhos Correlatos

Algumas técnicas de Data Mining analisam séries temporais que são observações consecutivas de uma variável durante intervalos de tempo [DUN 99]. Utilizando técnicas de regressão, análise de co-relação ou modelos de predição, dependências entre valores podem ser avaliadas no tempo, ou seja, verifica-se a existência de relação entre valores de um momento no tempo para valores em outro momento, num tempo passado ou futuro. Um survey sobre trabalhos na área pode ser encontrado em [KEO 02], onde diversos algoritmos foram testados em vários conjuntos diferentes de dados.

O trabalho de [SRI 96] trata de forma um pouco diferente o mesmo problema. Seu objetivo é encontrar padrões seqüenciais, assumindo que se tem como entrada um conjunto de seqüências, sendo cada seqüência uma lista de transações e cada transação um conjunto de itens. Isto permite descobrir implicações entre itens ou probabilidades condicionais, dada uma janela de tempo. Assim, seria possível descobrir que clientes que compram um produto X, voltam à empresa depois de 2 semanas para comprar o produto Y. Os autores sugerem a aplicação na área médica, assumindo que as seqüências podem ser de sintomas ou doenças e cada transação seria um conjunto de sintomas exibidos ou doenças diagnosticadas numa visita ou consulta. Assim, seria possível identificar os sintomas que precedem certas doenças.

Segundo [WON 00], um padrão seqüencial é uma série finita de elementos que se repete em um certo conjunto de dados. Seu trabalho combina técnicas de data mining

e de visualização para encontrar padrões em seqüências temporais. Ele se baseia no estudo da ordenação ou arranjo dos elementos, diferentemente da técnica de regra de associação, que estuda o quanto certos elementos ocorrem relacionados entre si. Dentro desse contexto, é possível analisar uma única série ou seqüência temporal para se encontrar padrões dentro da própria série ou seqüência, ou então fazer como [DUN 99], que compara várias séries temporais para tentar encontrar um padrão.

As técnicas de análise de séries temporais são mais apropriadas para valores numéricos e não sobre textos. No entanto, com o grande volume de textos disponíveis hoje em dia, surge a necessidade de analisar seqüências temporais de textos. Nesse sentido, há estudos que apresentam técnicas de Web Mining para analisar seqüências de páginas Web visitadas [SPI 00; SRI 00]. Nesse caso, as técnicas estatísticas são utilizadas para encontrar sub-seqüências que se repetem em diferentes sessões de visitas. Entretanto, a técnica considera somente a identificação da página e não seu conteúdo.

Já [FEL 98] avalia a distribuição de termos ou palavras-chave associadas a textos em diferentes grupos de textos com tempo associado. Isto permite descobrir quais termos tiveram maior ênfase em cada período de tempo e também descobrir tendências nas distribuições (crescentes ou decrescentes). O trabalho de [ROY 02] procura identificar tendências emergentes em coleções de textos. Para tanto, temas são extraídos dos textos e suas freqüências na coleção mapeadas em grupos de textos associados a intervalos de tempo. As tendências são identificadas como os temas que tiveram uma freqüência com crescimento acentuado durante um certo intervalo de tempo. Trabalhos similares podem ser encontrados em [POT 01; LAV 00; SWA 00; WON 00].

O método de análise de seqüências temporais discretas proposto neste artigo avalia a probabilidade condicional entre características dos textos ao longo do tempo, isto é, considerando uma determinada janela de tempo, mas considerando que há várias seqüências temporais e que os padrões devem existir significativamente nestas seqüências para serem considerados descobertas interessantes.

Uma contribuição do trabalho consiste nas características extraídas dos textos serem conceitos e não palavras ou termos. Os conceitos evitam o problema do vocabulário, permitindo analisar as dependências entre fenômenos do mundo real e não entre características do modelo de representação adotado. No caso dos prontuários, os conceitos representam sintomas, sinais, características pessoais e de comportamento dos pacientes e eventos da sua vida.

Uma outra contribuição consiste em os padrões interessantes não serem somente aqueles que possuem um suporte acima de um limiar, mas sim uma confiança significativa (probabilidade condicional). Portanto, o objetivo consiste em descobrir dependências entre características, ou seja, padrões nas seqüências que ocorrem não por coincidência mas por razões inerentes ao domínio. Isto permite saber se a presença de uma característica num texto pode condicionar a aparição de outra característica num texto seguinte (na mesma seqüência). No caso dos prontuários, pode-se avaliar se um remédio prescrito a um paciente num dado momento pode afetar o paciente em um momento futuro.

Finalmente, a dependência entre conceitos é avaliada em várias seqüências temporais. Assim, para que um padrão possa ser considerado relevante, ele deve aparecer em várias seqüências. No caso de prontuários médicos, isto significa que o

padrão deve aparecer num número significativo de pacientes e não somente na seqüência temporal correspondente a um único paciente.

3. Descrição do Método Utilizado

O processo de descoberta de conhecimento recebe como entrada diversas seqüências temporais, que são conjuntos de textos ordenados cronologicamente, tendo como objetivo encontrar correlações ou dependências entre conceitos presentes nos textos, de um texto para outro da mesma seqüência, mas avaliando a probabilidade condicional no conjunto de todas as seqüências temporais. Conceitos representam melhor que palavras o conteúdo dos textos. Isto porque as palavras apresentam problemas relacionados com o vocabulário, sinonímia, polissemia, variações léxicas, entre outros. A idéia central é realizar o processo de descoberta num nível acima das palavras, ou seja, não analisando somente os termos e expressões presentes nos textos, mas seu significado em relação aos fenômenos da realidade (pessoas, objetos, entidades, eventos e situações do mundo real). A abordagem baseada em conceitos para descoberta de conhecimento em textos foi testada com sucesso e apresentada em [LOH 01; LOH 03]. Portanto, conceitos permitem avaliar melhor a relação entre fenômenos e eventos do mundo real.

3.1. Identificação de conceitos em textos

Conceitos são extraídos dos textos por um método de classificação que analisa estatisticamente as palavras presentes nos textos. Este é feito de forma automática, desde que esteja disponível uma ontologia previamente definida. Essa ontologia é geralmente construída de forma semi-automática, ou seja, o processo é realizado por pessoas com auxílio de ferramentas de software.

Na ontologia estarão definidos os conceitos relevantes e as regras para sua identificação no texto. A definição de um conceito deve ser feita através de uma ou mais regras, nas quais devem ser indicados termos positivos e termos negativos. Todas as frases de todos os textos são comparadas com todas as regras de todos os conceitos. Para um conceito estar presente num texto, basta que uma regra seja verdadeira em relação a uma frase deste texto. Para uma regra ser verdadeira, todos os termos positivos da regra devem estar presentes na frase e nenhum termo negativo pode aparecer. Se uma das regras for verdadeira para a frase sendo analisada, então o conceito está presente na frase e, conseqüentemente, no texto. Por exemplo, o conceito “alcoolismo” poderia ser definido pelas regras: “álcool –nega” e “hálito etílico”. Nelas, o símbolo “–” indica um termo negativo e, nesse caso, o termo “nega” aparece para eliminar frases como “o paciente nega uso de álcool”.

Como resultado do processo tem-se associada a cada texto uma lista com os conceitos presentes nele. A qualidade desse método de identificação foi testada e apresentada em [LOH 01; LOH 03]. A taxa de erro na identificação de conceitos em prontuários textuais de psiquiatria ficou abaixo de 10%.

3.2. A técnica de análise de seqüência temporal sobre conceitos

A técnica proposta permite descobrir dependências entre conceitos que aparecem em textos dentro de uma mesma janela de tempo. O objetivo é saber se um conceito condiciona a aparição de outro no futuro. Os conceitos são identificados nos textos conforme o método descrito na seção anterior. Os textos a serem analisados no processo

de descoberta devem obrigatoriamente seguir uma ordem cronológica, formando uma seqüência temporal. Essas seqüências podem ser independentes, não havendo relação explícita entre textos de uma seqüência e de outra.

Para determinação das dependências entre conceitos é utilizada a probabilidade condicional, avaliada entre os conceitos de um texto em relação aos conceitos presentes nos textos seguintes da mesma seqüência. A técnica foi implementada analisando textos contíguos, ou seja, as dependências são avaliadas entre conceitos de um texto para seu sucessor na seqüência (do anterior para o seguinte). Dessa forma, a probabilidade de um conceito aparecer é avaliada em relação aos conceitos presentes no texto imediatamente anterior na seqüência temporal.

Assim, o resultado desta etapa é um conjunto de regras do tipo $X \rightarrow Y$, onde X e Y são conceitos, mas X aparece no texto imediatamente anterior ao texto onde Y aparece. Este método corresponde ao operador “immediately future” na lógica temporal de [ALL 84], indicando que Y vale no momento imediatamente seguinte ao em que X valeu. No caso, cada momento é representado por um texto na seqüência temporal. Cabe salientar que não há necessidade de os textos serem publicados em períodos de tempo regulares, bastando que eles formem uma seqüência ou ordem cronológica.

Série A: Conceitos presentes no primeiro texto: X, Y Conceitos presentes no segundo texto: W, Z Conceitos presentes no terceiro texto: K	Série C: Conceitos presentes no primeiro texto: X, Z Conceitos presentes no segundo texto: L
Série B: Conceitos presentes no primeiro texto: X, Z Conceitos presentes no segundo texto: W	

Figura 1: Exemplo de uma coleção com seqüências temporais de textos

A Figura 1 apresenta uma coleção hipotética contendo 3 séries temporais. A série A possui 3 textos, enquanto que as séries B e C possuem 2 cada. Um dos padrões que poderia ser descoberto nesta coleção é “ $X \rightarrow W$ ”, que pode ser interpretado como “W aparece imediatamente depois de X”. Entretanto, esta regra não acontece sempre; seu grau de confiança é de 66,66%, pois W aparece duas vezes após o X, mas, na seqüência C, X aparece num texto e W não aparece no texto seguinte. Também pode-se notar um padrão com menor confiança, que é “ $Z \rightarrow K$ ” (confiança de 33,33%).

A confiança de uma regra do tipo “ $X \rightarrow W$ ” é dada pelo número de vezes em que W aparece num texto imediatamente após o texto em que X aparece, dividido pelo número de vezes em que X aparece. A regra também possui um grau de suporte, que é dado pelo número de textos onde ocorrem os conceitos X e W em seguida.

4. Avaliação da Abordagem sobre Prontuários Médicos

A abordagem foi aplicada em prontuários médicos dos pacientes de uma clínica psiquiátrica. Eles contêm textos descrevendo a evolução do paciente durante sua estada na clínica. Cada evolução é registrada em um texto diferente, escrito de forma livre por um profissional da área. Os textos descrevem sintomas e sinais do paciente, resultados de exames psicológicos e físicos, eventos ocorridos e informações importantes (por exemplo, “o paciente diz ver bichos e ser perseguido”). As evoluções de cada paciente formam uma seqüência temporal. Não há uma regularidade de tempo entre as evoluções, entretanto, os textos respeitam a ordem cronológica em que foram escritos.

Acredita-se ser possível identificar padrões de comportamento e de métodos de tratamento nestas seqüências. Especialmente, tem-se interesse pela administração de

remédios. Supõe-se que seja possível identificar mudanças no comportamento ou estado do paciente após certos remédios terem sido prescritos pelos médicos e terem sido tomados pelos pacientes.

Cada texto está associado a um diagnóstico que representa a doença mental do paciente. Este diagnóstico foi decidido por um médico da clínica em um processo real e prévio. Entretanto, a indicação do diagnóstico não está explicitamente expressa no texto. A classificação usada para o diagnóstico segue as regras da Classificação Internacional de Doenças (CID-10) [CEN 89]. Foram usadas somente as classes de primeiro nível, as quais correspondem aos diagnósticos mais frequentes na clínica estudada, a saber: (a) transtornos mentais orgânicos (F00 a F09); (b) transtornos mentais e comportamentais devidos ao uso de substância psicoativa (F10 a F19); (c) esquizofrenia (F20 a F29); e (d) transtornos do humor ou transtornos afetivos (F30 a F39).

Foi utilizada uma ontologia com 97 conceitos. Destes, 65 eram referentes a características do paciente, tais como “agressividade”, “insônia”, “mora sozinho”, “tabagista”, e 32 referentes a remédios. Profissionais da área ajudaram na definição dos conceitos e respectivas regras de identificação. Foram coletados 1000 textos, correspondendo a evoluções de 30 pacientes. Cada texto estava associado a um único paciente e, portanto, pertencia a somente uma seqüência temporal. Além disto, cada texto era identificado pela sua ordem na seqüência temporal. Foram definidos como limiares mínimos: 20% de confiança e 2% de suporte.

4.1. Principais conhecimentos descobertos

Nessa seção serão discutidos os principais resultados encontrados. Eles são apresentados no formato de regras já descrito ($X \rightarrow Y$), significando que o aparecimento do conceito X num texto implica em ou determina a presença do conceito Y no texto seguinte. Para preservar marcas de remédios e substâncias ativas, os nomes foram alterados.

Um dos padrões encontrados em toda a coleção foi: “remédio X \rightarrow pensamento (conf = 66,67%)”, significando que “o uso de remédios do tipo X pode interferir no pensamento do paciente”. Também se pode interpretar tal descoberta como “2/3 dos pacientes que tomam remédios X têm o seu pensamento afetado num futuro breve, podendo levar a pensamentos mágicos ou de perseguição, por exemplo”.

A abordagem também foi aplicada em sub-coleções (seqüências temporais correspondentes a pacientes com um mesmo diagnóstico). Isto permitiu avaliar padrões por diagnóstico, uma vez que, por hipótese, pacientes com diagnósticos diferentes recebem tratamentos diferenciados e portanto devem apresentar padrões distintos. Analisando somente os pacientes com diagnóstico de esquizofrenia pode-se descobrir o padrão “remédio Y \rightarrow nervosismo (conf= 22,22%)”, significando que o uso do remédio Y pode induzir a sintomas de nervosismo em uma parcela pequena (22%).

No caso dos pacientes com diagnóstico de transtornos afetivos (incluindo a depressão), pôde-se notar os seguintes padrões: (a) remédio Z \rightarrow fala_{doente} (conf= 28,57%); (b) remédio Z \rightarrow nervosismo (conf=28,57%); (c) remédio Z \rightarrow melhora (conf = 28,57%); e (d) remédio W \rightarrow remédio K (conf=25,00%). O padrão (a) indica que o remédio do tipo Z acaba afetando negativamente a fala dos pacientes em 28% dos casos. O padrão (b) indica que, na mesma proporção, este remédio pode deixar o paciente mais nervoso em momentos posteriores. Mas também, em 28% dos casos, houve melhora

depois que o paciente tomou este remédio, conforme o padrão (c). O padrão (d) revela que 25% dos pacientes tiveram que tomar o remédio K após terem tomado o remédio W (num tempo posterior, por exemplo, após alguns dias).

O conhecimento descoberto foi avaliado por médicos especialistas em psiquiatria. Estes médicos estudaram os conhecimentos descobertos e concordaram que eram conhecimentos novos e úteis. Entretanto, as descobertas encontradas neste experimento podem ser consideradas hipóteses e merecem estudos mais aprofundados, os quais transcendem o escopo deste trabalho. O importante destes resultados é a demonstração da viabilidade da estratégia para descoberta de conhecimento (neste caso, na forma de hipóteses) em textos que obedeçam a seqüências temporais.

5. Conclusões

Neste trabalho foi apresentada uma abordagem para descoberta de conhecimento em textos que seguem uma ordem cronológica ou série temporal. A abordagem utiliza uma técnica de análise de seqüências temporais discretas. As características a serem analisadas são conceitos extraídos dos textos. Pelos experimentos realizados, foi possível verificar que a abordagem utilizada pode identificar dependências entre conceitos de um texto para outro, contíguos na mesma seqüência. Apesar de a coleção de textos não parecer muito volumosa, ela é significativa para o domínio em questão. Trinta pacientes formam uma parcela significativa dos internos da clínica. Além disto, isto não invalida a avaliação da técnica de análise de seqüência temporal, pois demonstrou-se ser possível descobrir padrões que não aparecem em coleções aleatórias. Como trabalhos futuros, planeja-se desenvolver ferramentas que implementem alguns operadores da lógica temporal de intervalos de [ALL 94], tais como X ocorreu antes de Y (*before*), durante Y (*during*) etc. Também estão previstas aplicações desta abordagem em textos de outros domínios, como por exemplo registros de ocorrências policiais.

Referências Bibliográficas

- [ALL 94] Allen, J.F. and Ferguson, G. Actions and Events in Interval Temporal Logic, *J. Logic and Computation* v.4, n.5, 1994.
- [CEN 89] Centro Brasileiro de Classificação de Doenças - Centro Colaborador da OMS para a Classificação de Doenças em Português, *Classificação Internacional de Doenças e de Problemas Relacionados a Saúde*, 10 ed., São Paulo: EDUSP, 1989.
- [DUN 99] Duncan, George; Gorr, Wilpen; Szczypula, Janusz, Forecasting analogous time series, In: Armstrong, S. (ed.), *Forecasting principles*, Kluwer Academic Publishers, 1999.
- [FEL 95] Feldman, Ronen; Dagan, Ido, Knowledge discovery in textual databases (KDT), In: *International Conference on Knowledge Discovery*, Montreal, 1995, p.112-117.
- [FEL 98] Feldman, Ronen; Dagan, Ido, Mining text using keyword distributions, *Journal of Intelligent Information Systems*, v.10, n.3, May/June, 1998, p.281-300.
- [KEO 02] Keogh, Eamonn; Kasetty, Shruti, On the need for time series data mining benchmarks: a survey and empirical demonstration, In: *ACM SIGKDD*, Edmonton, Canada, 2002, p.102-111.

- [LAV 00] Lavrenko, Victor et al., Mining of concurrent text and time series, In: *Workshop on Text Mining, ACM SIGKDD*, Boston, USA, 2000, p. 37-44.
- [LOH 01] Loh, S.; Oliveira, J.P.M.; Gastal, F.L., Knowledge discovery in textual documentation: qualitative and quantitative analyses, *Journal of Documentation*, v.57, n.5, September 2001, pp.577-590.
- [LOH 03] Loh, S.; Oliveira, J. P. M.; Gameiro, M. A., Knowledge discovery in texts for constructing decision support systems, *Applied Intelligence*, v. 18, n.3, 2003, pp.357-366 (special issue on Text and Web Mining).
- [MAL 00] Maltz, Michael D., Visualizing lives: new pathways for analyzing life course trajectories, *Journal of Quantitative Criminology*, v.16, n.2, 2000, p.255-281.
- [POT 01] Pottenger, William M.; Yang, Ting-hao, Detecting emerging concepts in textual data mining, in: Michael Berry (ed.), *Computational Information Retrieval*, SIAM, Philadelphia, August 2001.
- [ROY 02] Roy, Soma; Gevry, David; Pottenger, William M., Methodologies for trend detection in textual data mining, in: *Text mining Workshop, Second SIAM International Conference on Data Mining*, April, 2002.
- [SPI 00] Spiliopoulou, Myra et al., Improving the effectiveness of a Web site with Web usage mining, In: *Workshop WEBKDD '99*, Lecture Notes on Artificial Intelligence 1836, Springer-Verlag. July 2000.
- [SRI 96] Srikant, Ramakrishnan; Agrawal, Rakesh, Mining sequential patterns: generalizations and performance improvements, in: Peter M. G. Apers, Mokrane Bouzeghoub, Georges Gardarin (Eds.), *Advances in Database Technology - EDBT'96*, International Conference on Extending Database Technology, 5, Avignon, France, March 25-29, 1996, Lecture Notes in Computer Science. v.1057, Springer-Verlag, Springer 1996, p.3-17.
- [SRI 00] Srivastava, Jaideep et al., Web usage mining: Discovery and applications of usage patterns from Web data, *ACM SIGKDD Explorations*, v.1, n.2, January 2000.
- [SWA 00] Swan, Russell; Jensen, David, TimeMines: constructing timelines with statistical models of word usage, In: *Workshop on Text Mining, ACM SIGKDD*, Boston, USA, 2000.
- [TAN 99] Tan, Ah-Hwee, Text mining: the state of the art and the challenges, In: *Pacific-Asia Workshop on Knowledge Discovery from Advanced Databases - PAKDD '99*, Beijing, 1999, p. 65-70.
- [WON 00] Wong, P. et al., Visualizing sequential patterns for text mining, In: *IEEE Symposium on Information Visualization*, Salt Lake City, Utah, October 2000, p.105-111.
- [MEI 05] Mei, Q , Zhai, C. Discovering evolutionary theme patterns from text: an exploration of temporal text mining, In: *ACM SIGKDD 2005*, August 21–24, 2005, Chicago, Illinois, USA. p.198-207.