em questão
E-ISSN 1808-5245

Descoberta de conhecimento a partir de informações não estruturadas por meio de técnicas de correlação e associação

Marina Carradore Sérgio
Doutoranda; Universidade Federal de Santa Catarina (UFSC), Florianópolis, SC, Brasil;
marinacarradore@egc.ufsc.br

Thales do Nascimento da Silva
Doutorando; Universidade Federal de Santa Catarina (UFSC), Florianópolis, SC, Brasil;
thales788@gmail.com

Alexandre Leopoldo Gonçalves
Doutor; Universidade Federal de Santa Catarina (UFSC), Florianópolis, SC, Brasil;
a.l.goncalves@ufsc.br

Resumo: O atual momento da tecnologia vem promovendo meios para o aumento exponencial no volume de informações disponíveis na internet ou em organizações. Considerando que grande parte desta informação encontra-se em formato textual, este fato representa um desafio para as áreas de coleta, armazenamento, recuperação e análise de informações visando à explicitação de conhecimento. Este trabalho tem como objetivo apresentar um modelo para Descoberta de Conhecimento com base nas técnicas de correlação e associação temporal a partir de grandes coleções de documentos. Os procedimentos metodológicos utilizados compreenderam uma pesquisa descritiva e exploratória, envolvendo artigos coletados da base de dados *Science Direct*® como uma ferramenta para a coleta e a análise dos dados. Através deste tipo de informação é possível extrair regras, padrões, tendências e redes, capazes de auxiliar no processo de tomada de decisão nas organizações a fim de gerar vantagem competitiva. Como principal contribuição destaca-se a proposição de um modelo voltado ao entendimento de aspectos temporais, considerando relacionamentos factuais (através de correlações) ou não (através de associação) entre termos de um domínio.

Palavras-chave*:* Descoberta de conhecimento. Correlação. Associação. Informações não estruturadas. Temporalidade.

em questão
E-ISSN 1808-5245

1 Introdução

As evoluções dos meios computacionais juntamente com o aumento da capacidade de processamento, armazenamento e conectividade, estão provocando um crescimento exponencial no volume de informação (FLEUREN; ALKEMA, 2015). Pesquisas realizadas por Hilbert e López (2011) concluíram que até 2007 haviam sido produzidos 295 exabytes de informações, e segundo Wu et al. (2014) todos os dias 2,5 quintilhões de bytes de dados são criados, sendo que 90% dos dados produzidos no mundo foram gerados nos últimos anos. Estima-se que até 2020 o volume de informação, a nível mundial, cresça em 35 trilhões de gigabytes (GANTZ; REINSEL, 2010).

Aproximadamente 80% destas informações se encontram em formato textual (SOMASUNDARAM; SHRIVASTAVA, 2011; RÊGO, 2013). Este cenário promove desafios quanto à coleta, armazenamento, recuperação e análise de informação não estruturada a ponto de gerar conhecimento, com o intuito de servir como uma fonte de vantagem competitiva para as organizações. Para lidar com tais desafios tornam-se necessários modelos, processos, metodologias, entre outros, para identificar e reaproveitar conhecimentos. Entre estes se encontra o processo de Descoberta de Conhecimento em Texto (do inglês, *Knowledge Discovery in Text* - KDT) entendido como uma versão da Descoberta de Conhecimento em Bases de Dados (do inglês, *Knowledge Discovery in Database* - KDD) voltada à manipulação de informação não estruturada. Considerando o extenso volume de documentos disposto em linguagem natural, o processo de KDT tornou-se o foco de diversos estudos (HASHIMI; HAFEZ; MATHKOUR, 2015). Este processo tem como objetivo desvendar padrões e tendências, classificando e comparando os mais variados documentos.

Em razão de sua potencialidade, torna-se de suma importância o desenvolvimento de modelos embasados em técnicas que possibilitem simplificar o processo de descoberta de padrões em bases dessa natureza. Dentre as técnicas apresentadas na literatura encontram-se a Correlação e Associação. A Correlação é responsável por determinar o grau de relacionamento entre duas

variáveis, enquanto que a Associação se encarrega de evidenciar relacionamentos indiretos, buscando explicitar conexões potencialmente úteis entre os termos.

Afim de revelar padrões latentes em grandes coleções de documentos que estejam disponíveis no meio da web ou em organizações ao mesmo passo que envolvidos em um determinado período temporal, o desenvolvimento desta pesquisa se dá por meio da motivação para prover soluções para os desafios de produzir conhecimento útil ao processo de tomada de decisão pelas organizações. Metodologias, modelos, técnicas e algoritmos provenientes de diferentes áreas que promovam suporte à Descoberta de Conhecimento são fundamentais para o desenvolvimento de sistemas capazes de lidar com tais demandas.

Os sistemas atuais de descoberta de conhecimento em bases de dados não estruturados aplicam métodos de correlação/associação objetivando extrair conhecimento, porém não fornecem todos os elementos dos processos de armazenamento, pré-processamento e recuperação do conteúdo textual considerando a dimensão temporal. Deste modo, o presente artigo visa propor um modelo de descoberta de conhecimento aplicado a bases de documentos textuais por meio de técnicas de correlação e associação de maneira temporal com suporte da computação distribuída.

As demais seções do artigo são estruturadas de modo que a seção 2 apresente os principais referenciais teóricos envolvidos na proposição deste artigo, enquanto a seção 3 apresenta o modelo proposto e na seção 4 é explanado sobre a metodologia de pesquisa utilizada para o desenvolvimento deste trabalho. A discussão dos resultados ocorre ao longo da seção 5, até que finalmente sejam detalhadas as considerações finais e os trabalhos futuros na última seção.

2 Descoberta de conhecimento

Os processos de Descoberta de Conhecimento se destinam à análise de grandes conjuntos de dados (FENG, 2010), buscando padrões que resultem em

conhecimento útil e que tenham surgido como uma solução fundamental para a compreensão do valor real dos dados coletados, objetivando auxiliar o processo de tomada de decisão nas organizações.

2.1 Descoberta de conhecimento em bases de dados

O processo de Descoberta de Conhecimento em Bases de Dados tem por objetivo identificar e desvendar relacionamentos implícitos entre as informações armazenadas nas bases de dados organizacionais (SILVA; ROVER, 2011). O KDD surgiu como uma alternativa para solucionar o problema da sobrecarga de dados na era da informação digital (SHABBIR et al., 2014). Para Zhu et al. (2013), o processo de KDD se constitui na análise e na exploração automática ou semiautomática de grandes volumes de dados, objetivando desvendar regras e padrões significativos. Os padrões, após descobertos, são empregados no auxílio à tomada de decisão em determinado contexto (CAO et al., 2010). Desta forma, o processo visa a descoberta de conhecimento interessante e útil (VASHISHTHA; KUMAR; RATNOO, 2012), e o KDD destina-se a facilitar e acelerar a extração de conhecimento a partir de fontes de dados persistentes (NOACK; SCHMITT, 2013).

O processo de descoberta de conhecimento em bases de dados compreende as etapas de seleção dos dados, o pré-processamento que adequa os dados aos algoritmos, a mineração efetiva dos dados que compreendem o uso de técnicas geralmente baseadas na Inteligência Artificial ou Estatística (MAIA; SOUZA, 2010), a validação dos resultados e a análise e interpretação dos resultados para aquisição do conhecimento. O principal objetivo deste processo é a tradução de dados brutos em informações relevantes para posterior utilização e descoberta (ZHU et al., 2013).

2.2 Descoberta de conhecimento em textos

O processo de Descoberta de Conhecimento em Textos (KDT) assemelha-se ao

KDD, porém é voltado ao tratamento de documentos textuais (HASHIMI; HAFEZ; MATHKOUR, 2015). Tanto o KDT quanto o KDD referem-se ao processo de extração de padrões não triviais e de conhecimento útil (ZHU et al., 2013; HASHIMI; HAFEZ; MATHKOUR, 2015). Entretanto, a área que envolve o KDT torna-se mais complexa devido à falta de estruturação da informação descrita em linguagem natural (ZHU et al., 2013).

Documentos textuais possuem uma estrutura que necessita da aplicação de técnicas especializadas para serem analisados por sistemas computacionais, devido ao significado implícito atribuído a cada palavra na linguagem humana (SABOL et al., 2009). Segundo Hashimi, Hafez e Mathkour (2015), grande parte do conhecimento existente está disposto no formato textual, e em função deste motivo, tal conhecimento precisa ser identificado, representado e manipulado de modo a tornar-se potencialmente útil para as organizações.

2.3 Estrutura de apresentação da informação

A estrutura de apresentação da informação pode ser dividida em estruturada, semiestruturada e não estruturada. A informação estruturada é representada normalmente em tabelas, gerenciadas por softwares de banco de dados (RAMOS; BRÄSCHER, 2009). A informação semiestruturada, por sua vez, é normalmente apresentada entre marcadores (*tags*), tais como documentos XML e HTML (CHEN et al., 2009), onde a estrutura de apresentação possibilita o entendimento por parte dos meios computacionais. Por outro lado, a informação não estruturada é disposta em linguagem natural e não segue um padrão de apresentação (LIM; LIU; LEE, 2009), ou seja, não contém estrutura tabular e nem marcação. É o caso de exemplos como e-mails, artigos, comentários em redes sociais e documentos na Web.

2.4 Modelos baseados em coocorrência

No processo de Descoberta de Conhecimento e considerando fontes de

informação não estruturadas, torna-se necessário o emprego de técnicas para realizar a agregação da informação. Os modelos baseados em coocorrência entre termos[1] permitem evidenciar a combinação destes considerando um conjunto de dados.

O termo "correlação" significa literalmente "correlacionamento", sendo possível evidenciar o grau de relacionamento entre duas variáveis. O grau de correlação entre os termos contidos nos documentos textuais pode ser representado em cálculos oriundos da estatística. A finalidade do cálculo de correlação é a determinação da força do relacionamento entre dois elementos em análise (BARALIS et al., 2013). Entre os modelos utilizados para determinar a correlação encontram-se a Frequência Conjunta, Média e Variância, Teste T (MANNING; SCHÜTZE, 1999), o Chi-square (CHURCH; MERCER, 1993), o Phi-squared (CHURCH; GALE, 1991; CONRAD; UTT, 1994) e a Informação Mútua (CHURCH; HANKS, 1990).

Neste trabalho o cálculo utilizado foi o Phi-squared, que segundo Church e Gale (1991) é definido como:

$$\Phi^2 = \frac{(ad - bc)^2}{(a+b)(c+d)(a+c)(b+d)}, onde\ 0 \leq \Phi^2 \leq 1$$

A aplicação do Phi-squared utiliza uma tabela 2*2 (tabela de contingência), conforme observado no Quadro 1.

Quadro 1 - Tabela de contingência.

	t_2	$\overline{t_2}$
t_1	a	b
$\overline{t_1}$	c	d

Fonte: Sérgio (2013)

Sendo que *a* representa a frequência em que os termos t_1 e t_2 ocorrem de forma conjunta, *b* representa as ocorrências do termo t_1 onde não há a presença de t_2, *c* representa as ocorrências de t_2 sem a presença de t_1, e *d* é o tamanho da coleção de documentos menos o número de documentos que não contenham t_1 e/ou t_2, sendo *d=N-a-b-c*, onde *N* é o tamanho da base.

2.5 Associação entre termos

O tópico anterior apresenta meios para o estabelecimento de relacionamentos diretos entre termos. Ainda que isto possa promover uma visão inicial do contexto em que os termos estejam inseridos, relacionamentos diretos não são capazes de capturar a dinâmica de associação entre os termos que podem promover um entendimento mais detalhado sobre determinado domínio de análise. Neste sentido, o processo de associação é responsável por evidenciar relacionamentos indiretos, com o objetivo de explicitar conexões potencialmente úteis entre os termos.

A área biomédica e da bioinformática vem provocando grandes avanços envolvendo a associação entre termos, tendo em vista revelar novos conhecimentos (WOSZEZENKI; GONÇALVES, 2013). Na base destas pesquisas encontram-se os trabalhos relativos à área de Descoberta Baseada em Literatura (DBL – do inglês *Literature-Based Discovery*), proposta inicialmente por um cientista norte americano, Don R. Swanson, que efetuou pesquisas envolvendo a área biomédica e a descoberta de relacionamento implícito entre padrões (SWANSON, 1986). O objetivo principal da Descoberta Baseada em Literatura é desvendar relacionamentos implícitos em bases científicas, com o intuito de originar potenciais proposições para novas descobertas (SMALHEISER, 2012).

O modelo vetorial no contexto de associação indireta visa determinar o coeficiente de semelhança entre um conjunto de termos. Cada termo a ser analisado possui o seu vetor de contexto determinado pelas relações estabelecidas através de correlação. O vetor de contexto é responsável por descrever determinado termo em que cada posição é preenchida com um termo relacionado e o seu grau de correlação. Para que se obtenha a similaridade desejada são utilizadas algumas medidas, sendo que medidas como o índice Jaccard, o índice Dice, a medida Overlap (máxima e mínima), a medida do Cosseno e a medida do Pseudo-cosseno (JONES; FURNAS, 1987; EGGHE; MICHEL, 2002) recebem destaque. Neste trabalho considerou-se a equação do

Cosseno definida da maneira como se segue:

$$\cos\theta = \frac{\sum_{i=1}^{n}(w_{qi} \times w_{oi})}{\sqrt{\sum_{k=1}^{n}(w_{qk})^2} \times \sqrt{\sum_{j=1}^{n}(w_{oj})^2}}$$

Onde w_{qi} e w_{qk} representam os pesos dos *i*th e *k*th termos do vetor *q*, e w_{oi} e w_{oj} representam os pesos dos *i*th e *j*th termos do vetor *o*.

2.6 Computação distribuída como suporte ao extenso volume de dados

Entre os anos de 1945 e 1985, computadores ocupavam um grande espaço e tinham um custo elevado, e, de modo geral, estes computadores trabalhavam de forma independente devido à inexistência de uma forma de interligá-los (TANENBAUM; STEEN, 2008). Com a evolução das redes de computadores, tornou-se possível a conexão de computadores, e progressivamente a velocidade atingida nestas conexões tornava-se cada vez maior.

A partir do cenário descrito e uma necessidade de maior capacidade de processamento, a computação distribuída destaca-se como uma alternativa viável a esta demanda. O conceito de "computação distribuída" pode ser definido como um sistema composto por vários componentes de hardware ou software que se comunicam, compartilham recursos e coordenam suas ações por meio da troca de mensagens (COULOURIS et al., 2013). Tanenbaum e Steen (2008) citam algumas vantagens de sistemas distribuídos quando comparados a sistemas centralizados:

a) maior poder de processamento: um sistema distribuído pode ter mais capacidade de processamento em relação a servidores centralizados;
b) crescimento incremental: o poder computacional pode crescer incrementalmente;
c) compartilhamento de dados e recursos: tornam-se possíveis

aplicações envolverem máquinas separadas geograficamente;

d) maior confiabilidade: o sistema pode continuar funcionando mesmo perdendo alguns componentes;

e) menor custo/benefício: os sistemas distribuídos têm melhor custo/beneficio em relação aos sistemas centralizados.

3 Modelo proposto

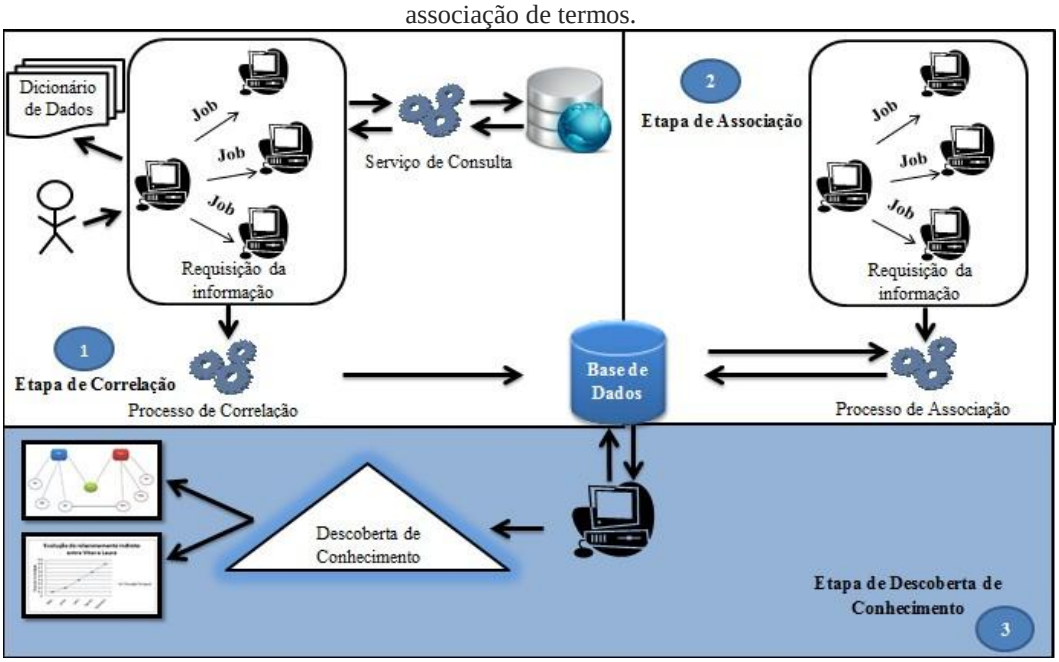
Aliada ao fácil acesso aos meios de comunicação, a rápida evolução dos meios de armazenamento levou a um veloz aumento no volume de informação. Grande parte da informação produzida encontra-se em formatos não estruturados, como textos em geral. E este tipo de informação, por não possuir uma estrutura formal, torna-se difícil de ser analisada.

Com a utilização do processo de Descoberta de Conhecimento em Textos (KDT), é possível extrair conhecimento desta fonte de informação. Apesar disso, a aplicação do KDT não é trivial, principalmente devido ao grande volume e ao fator de ambiguidade existente na informação. Outro aspecto de fundamental importância é o fator temporal, característica que permite descobrir comportamentos que descrevam fatos que já ocorreram ou que podem vir a ocorrer. O fator temporal é apontado como uma limitação adicional das abordagens existentes, pois estes trabalhos tentam determinar conexões implícitas significativas, considerando a distribuição dos termos ou conceitos de um corpus em um único ponto do tempo (COHEN; SCHVANEVELDT, 2010). A detecção de associações presentes em um conjunto de análise num espaço temporal pode ser vista como a previsão de futuras ligações explícitas (COHEN; SCHVANEVELDT; WIDDOWS, 2010; YETISGEN-YILDIZ; PRATT, 2009). Mudanças no grau de associação ao longo do tempo seriam importantes para prever conexões explícitas no futuro.

O presente artigo propõe um modelo computacional utilizando como base as técnicas de Correlação e Associação entre termos e a computação distribuída para a descoberta de conhecimento, com destaque para a dimensão

temporal. A Figura 1 ilustra as etapas que compõem o modelo proposto. Tais etapas possibilitam a interconexão do conteúdo textual representado por conceitos em um domínio de análise, cujo objetivo é prover suporte ao processo de descoberta de conhecimento.

Figura 1 - Modelo proposto de descoberta de conhecimento com base na correlação e associação de termos.

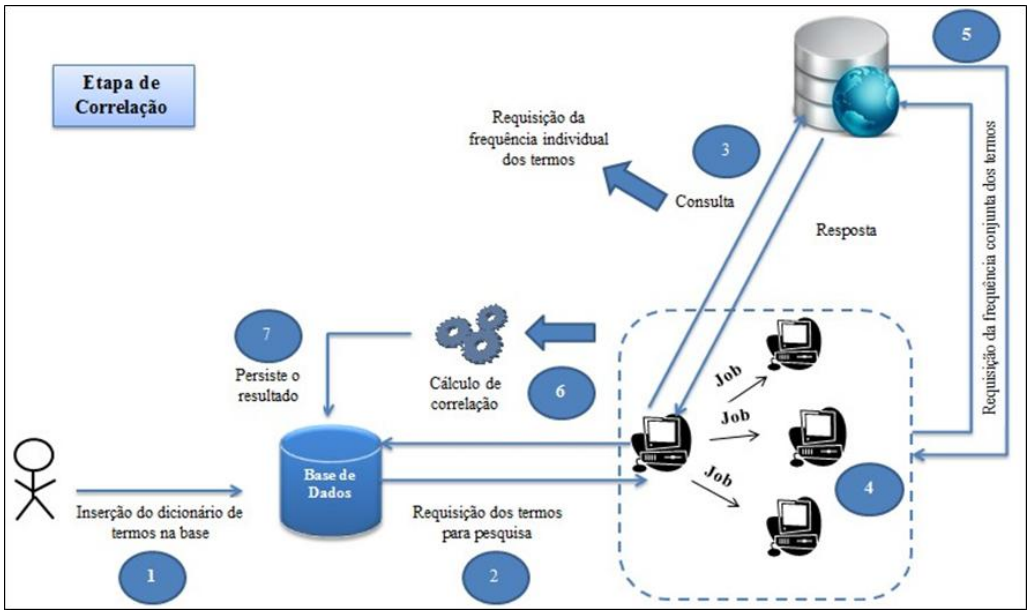


Fonte: Sérgio (2013)

3.1 Etapa de correlação

A etapa 1 é detalhada a partir da Figura 2. O primeiro passo consiste na definição de um dicionário que contenha os termos necessários para uma determinada análise, juntamente com a indexação de todos os documentos coletados para compor a base de dados. Neste ponto torna-se fundamental a intervenção de um especialista de domínio, pois o mesmo é responsável por definir e registar na base de dados os termos que serão utilizados no processo de correlação, e posteriormente no processo de associação.

Figura 2 – Detalhamento da etapa de correlação.



Fonte: Sérgio (2013)

No passo 2, a aplicação requisita uma lista de termos, sendo que estes foram inseridos na etapa anterior. No passo seguinte, levando em consideração cada termo a constar na lista, através de um serviço de consulta verifica-se em quais documentos o termo é mencionado. Ao final deste passo, obtém-se uma lista de termos e suas respectivas frequências individuais representando o número de documentos que contenham o termo pesquisado.

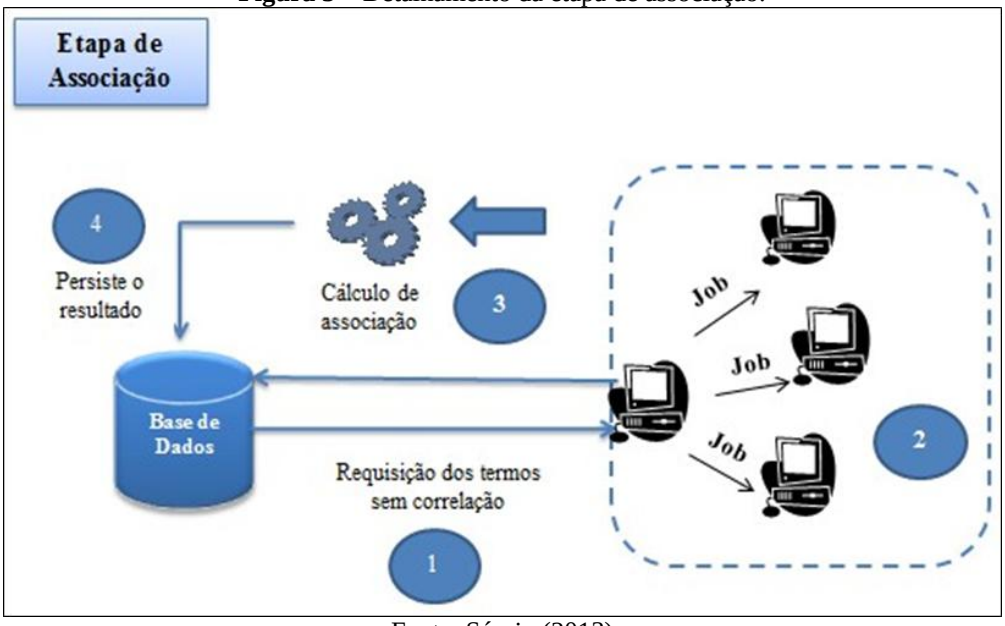
No passo 4, monta-se uma estrutura que seja capaz de prover todos os dados, de modo que o cálculo de correlação possa ser executado distribuidamente. Para que as tarefas fossem executadas de forma distribuída, foi utilizado o *framework/middleware* GridGain®. O GridGain® possibilita o desenvolvimento de aplicações distribuídas de alto desempenho e escalabilidade (IVANOV; DMITRIY, 2012). Já no passo 5, cada nodo que integra a rede distribuída é responsável por obter a frequência conjunta dos termos. Em outras palavras, é requisitado ao servidor de consulta a quantidade de documentos em que dois termos quaisquer aparecem conjuntamente. Para realizar este passo, todo nodo possui um termo origem e uma lista de termos destino. Sendo assim, o nodo calcula a frequência conjunta do termo origem com cada termo destino que compõe a lista. Já tendo obtido os valores da frequência individual e da frequência conjunta, a partir do passo 6 é possível calcular o coeficiente de

correlação que representa a força de correlação entre dois termos. O cálculo utilizado foi apresentado anteriormente no item sobre Modelos baseados em coocorrência. Finalmente, ao chegar no último passo, cada nodo da rede é responsável por persistir os coeficientes de correlação que calculou. A quantidade de nodos criados é igual o tamanho da lista de termos menos um.

3.2 Etapa de associação

O primeiro passo desta etapa é a requisição dos termos sem correlação. Sendo assim, a etapa é realizada apenas com termos que não possuem correlação direta, uma vez que a partir dos dados gerados objetivam-se análises onde exista uma tendência de aumento no coeficiente de associação ao longo do tempo, no entanto antes que estes passem a coocorrerem em um mesmo documento. A Figura 3 detalha a etapa de associação.

Figura 3 – Detalhamento da etapa de associação.



Fonte: Sérgio (2013)

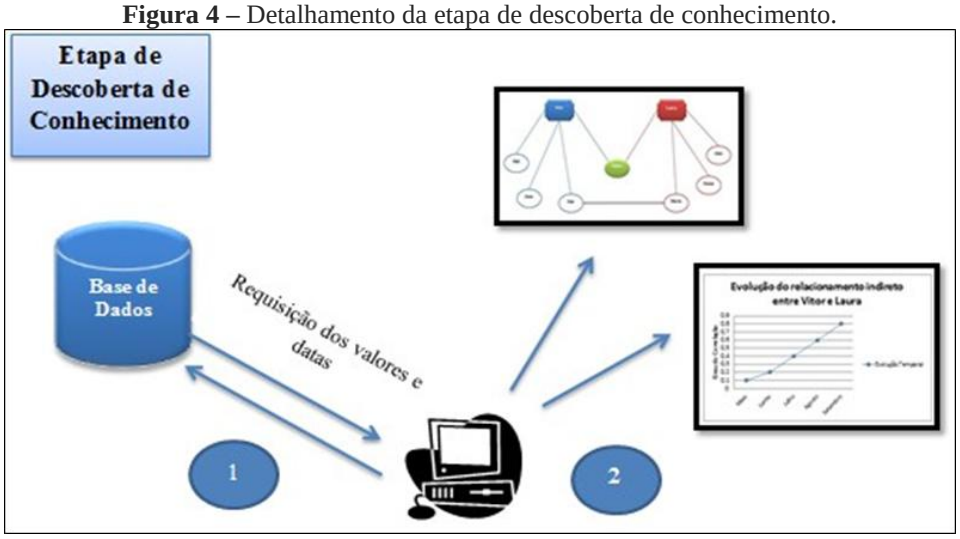
Em seguida realiza-se a divisão dos nodos da rede distribuída que serão responsáveis por calcular o coeficiente de associação. Esta divisão tem como objetivo verificar a associação de uma lista de termos em diferentes datas, de modo que seja possível que se aplique uma análise de associação temporal. É no

passo seguinte que o coeficiente de associação é calculado por meio da utilização do modelo vetorial apresentado anteriormente, de modo que no último processo relativo à esta etapa, cada nodo realize a inserção do resultado do cálculo de associação na base de dados.

3.3 Etapa de descoberta de conhecimento

Após as etapas de associação e correlação serem concluídas pelos nodos da rede, obtém-se o grau de similaridade entre os termos da pesquisa. Neste momento é realizada a descoberta de conhecimento subsidiada pelos passos anteriores. A base de dados resultante permite que seu conteúdo seja explorado, visando à obtenção de tendências e padrões (HASHIMI; HAFEZ; MATHKOUR, 2015) que auxiliem na descoberta de conhecimento relevante e útil para o apoio a tomada de decisão. Este conhecimento pode ser exposto através de gráficos de correlação e associação, histogramas, e mesmo mapas de tópicos, temporais ou não. Dentre as possibilidades citadas, a característica da análise temporal possibilita um acompanhamento da possível evolução do grau de associação entre dois termos. A Figura 4 ilustra o processo de descoberta de conhecimento.

Figura 4 – Detalhamento da etapa de descoberta de conhecimento.



Fonte: Sérgio (2013)

4 Metodologia da pesquisa

Os procedimentos metodológicos utilizados neste trabalho são de natureza descritiva e exploratória, utilizando a base de dados *Science Direct*® como fonte de coleta dos artigos para a análise dos dados.

4.1 Detalhamento do cenário de pesquisa

A construção do cenário de aplicação envolveu a coleta de artigos na base de dados *Science Direct*®, com o objetivo de evidenciar relações temporais existentes entre termos de determinado domínio. Critérios como a abrangência de áreas, assim como a confiabilidade e credibilidade, o volume de artigos publicados (aproximadamente 12 milhões) e os filtros de pesquisa se fizeram decisivos para escolha da base de coleta.

Abaixo, o Quadro 2 expõe os elementos da pesquisa para criação da base de dados. Na primeira e segunda coluna são apresentados os termos pesquisados e o foco da análise. Na terceira coluna, é apontado o período de realização da coleta e na última coluna, o momento em que ocorre a coocorrência entre os termos em pelo menos um documento.

Quadro 2 - Elementos de pesquisa para montagem da base de dados.

Termo de Pesquisa 1	Termo de Pesquisa 2	Período de realização da pesquisa	Momento em que ocorre a coocorrência
Biotechnology	*Genetic Engineering*	1993 a 2002	2003
Nanotechnology	*Medicine*	1984 a 1993	1994

Fonte: Sérgio (2013)

O período de coleta e extração dos dados para compor a base de dados compreendeu o ano de 2013. Foram coletados 313 artigos para o primeiro estudo de caso e 238 artigos para o segundo, conferindo um total de 551 documentos. No primeiro estudo de caso, os termos de consulta foram:

("Biotechnology" and "Genetic Engineering"), enquanto que no segundo os termos analisados foram (*"Nanotechnology" and "Medicine"*). A escolha dos termos foi motivada pela presença de termos relacionados à área da saúde em pesquisas envolvendo modelos de Descoberta de Conhecimento. A pesquisa foi realizada considerando a presença do termo no documento como um todo.

Para o primeiro estudo de caso *("Biotechnology" and "Genetic Engineering")*, compreendendo apenas o período de associação – ou seja, quando não existe coocorrência –, foram coletados 138 documentos. No período em que passa a existir coocorrência e a determinação da correlação se torna possível, foram coletados 175. Para o segundo estudo de caso *(Nanotechnology and Medicine), coletou-se* 185 documentos para o período de associação, e 53 para o período de correlação. Para criação da base de dados foi necessário extrair as meta-informações dos artigos e estruturá-los na forma de documentos XML – processo que foi realizado manualmente. E em razão do processo de coleta manual, optou-se por selecionar apenas uma quantidade limitada de documentos por ano.

O documento XML criado é composto por um identificador sequencial, o título, o ano, o nome do(s) autor(es) com sua(s) respectiva(s) organização(ões) e as palavras-chave. Caso as palavras-chave não existissem o documento era lido e as palavras-chave elencadas no arquivo XML correspondente. Na etapa seguinte os arquivos no formato XML e os documentos completos em formato PDF foram indexados via um servidor de indexação visando permitir as consultas para a identificação das frequências individuais e conjuntas dos termos. Considerando as palavras-chave dos artigos obteve-se 710 termos que foram utilizados no primeiro estudo envolvendo *Biotechnology* e *Genetic Engineering*, e 506 termos utilizados no segundo estudo envolvendo *Nanotechnology* e *Medicine*.

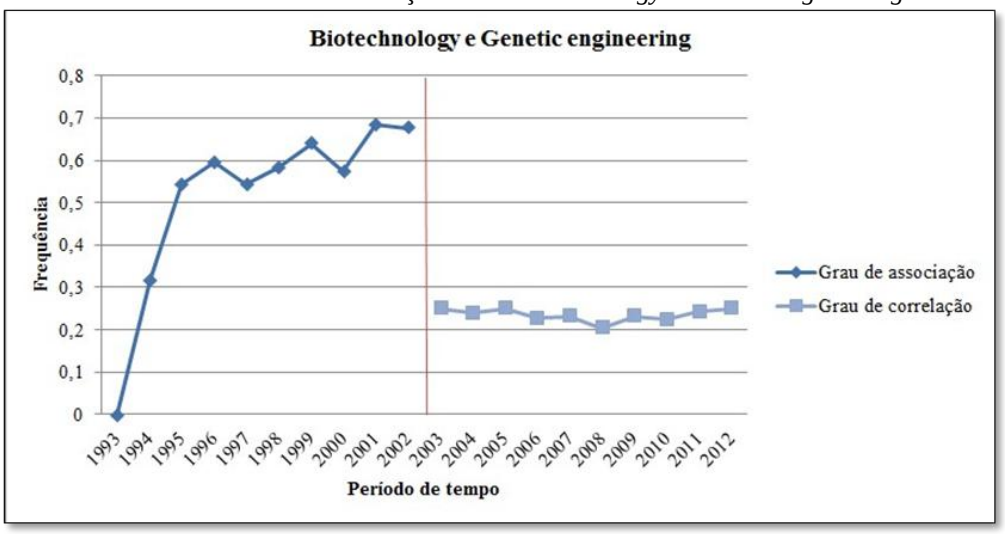
5 Discussão dos resultados

O modelo apresentado neste artigo promove suporte à evolução temporal dos relacionamentos entre termos. Por meio de gráficos e mapas de tópicos, com o objetivo de explicitar conhecimento em bases textuais e consequentemente auxiliar na tomada de decisão, se busca demonstrar a evolução dos relacionamentos entre termos.

Baseado nas frequências tanto individuais quanto conjuntas dos termos pesquisados é possível aplicar o cálculo da equação *Phi-squared* para que se obtenha o grau de correlação. Posteriormente, a partir da correlação aplica-se o cálculo do modelo vetorial para a obtenção do grau de associação entre dois termos.

Ao se utilizar os resultados como base, torna-se possível gerar análises sobre os dados observados. Tais análises podem indicar tendências associativas e apontar possíveis correlações passíveis de investigação em determinado período. O Gráfico 1 apresenta o momento em que ocorre a associação entre os elementos em análise e o momento posterior, a correlação. Neste caso, pode-se observar que o grau de associação entre *Biotechnology* e *Genetic Engineering* evolui até 2002, ou seja, compartilham termos presentes na representação vetorial, porém não coocorrem em um mesmo documento. E a partir de 2003, os termos passam a ser mencionados conjuntamente.

Gráfico 1 - Grau de associação entre *Biotechnology* e *Genetic engineering*

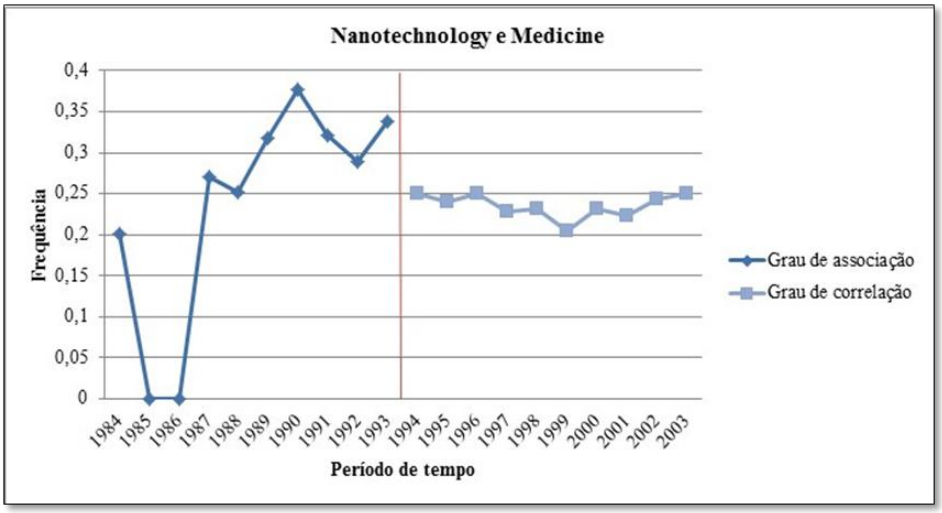


Fonte: Sérgio (2013)

O gráfico acima apresenta uma evolução do comportamento associativo entre dois termos quaisquer de interesse em uma análise. Deste modo, o aumento da tendência pode disparar alertas, visando uma análise mais detalhada nas fontes de informação. Cabe mencionar que a evolução da associação não garante que coocorrências irão acontecer de fato, mas criam indícios que podem auxiliar na tomada de decisão. Este tipo de gráfico tem impacto nos mais variados domínios, sejam científicos ou mesmo análises de contexto social.

Por outro lado, a evolução do grau de associação entre os elementos pode não ocorrer de maneira incremental. Entretanto, tal comportamento não indica que os termos não possam ser mencionados conjuntamente, visto que a determinação da coocorrência pode acontecer ao acaso. No Gráfico 2 pode-se observar um cenário em que a associação não é crescente, mas conduz a coocorrência mesmo assim.

Gráfico 2 – Evolução temporal entre os termos *Nanotechnology* e *Medicine*



Fonte: Sérgio (2013)

Os resultados obtidos na segunda análise, ainda que modestos, apontam um incremento nos valores, que vão de 0,193 para 0,329, o que equivale a aproximadamente 70% na evolução do grau de associação.

Além do comportamento apresentando nas análises acima, a associação entre dois termos pode sofrer decréscimos ao longo do tempo, indicando um afastamento dos mesmos. Este afastamento poderia, por exemplo, ser explicado pela evolução de determinada área em que um termo ou passa a ter menos

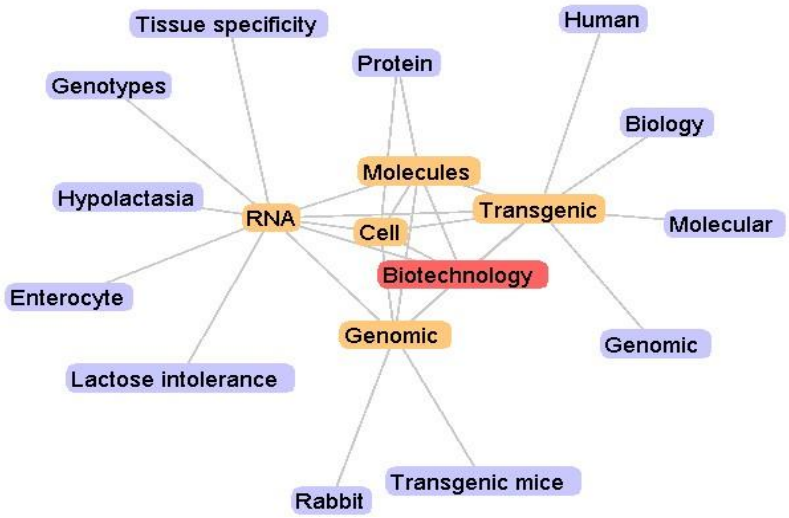
influência ou se altera para termos correlatos, uma vez que a designação original não representava adequadamente o conceito.

Os dados obtidos contribuem ainda para a geração de novos conhecimentos, possibilitando a exploração de conteúdo centrado na obtenção de padrões e tendências que possam conduzir a descoberta de conhecimento através de ferramentas que considerem os aspectos visuais de como os termos se interconectam.

Os mapas de tópicos estão entre as ferramentas que possibilitam este tipo de exploração, objetivando auxiliar no entendimento de determinado domínio de análise. A escolha de mapas de tópicos como meio de representação foi motivada pel a sua utilidade quanto à representação e descrição da informação, e bem como a estrutura conceitual de determinado domínio (AHMED; MOORE, 2005).

A elaboração dos mapas de tópicos é conduzida recursivamente, selecionando em cada nível do mapa os cinco termos mais relacionados a determinado termo de interesse – neste caso, *Biotechnology*, *Genetic Engineering*, *Nanotechnology* e *Medicine*. A expansão de cada mapa considerou dois níveis a partir do termo central. A Figura 5 apresenta o mapa de tópicos gerado a partir do termo *Biotechnology*.

Figura 5 – Mapa de tópicos referente ao termo *Biotechnology*

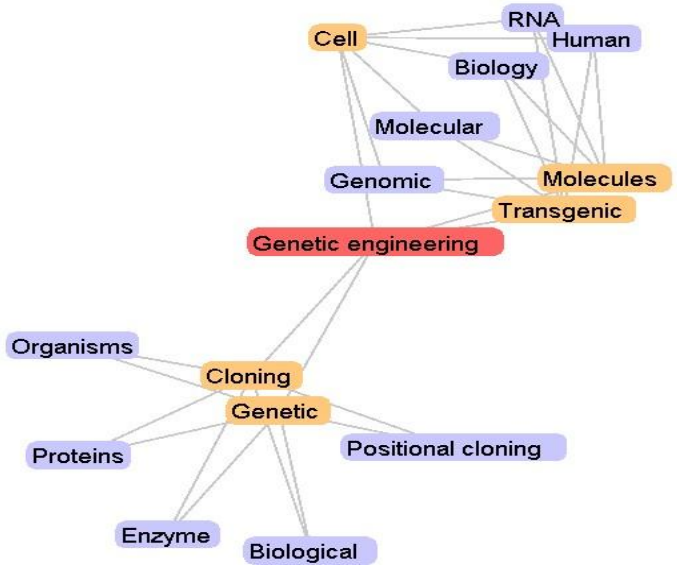


Fonte: Sérgio (2013)

As ligações entre os termos não possuem direção. Como mencionado, cada termo a partir da origem (termo de interesse) se conecta aos demais termos

em que possui a maior correlação. A cor vermelha representa o termo de análise, a cor amarela representa o primeiro nível, e a cor azul o segundo nível. A Figura 6 apresenta o mapa de tópicos gerado a partir do termo *Genetic Engineering*.

Figura 6 – Mapa de tópicos referente ao termo *Genetic Engineering*

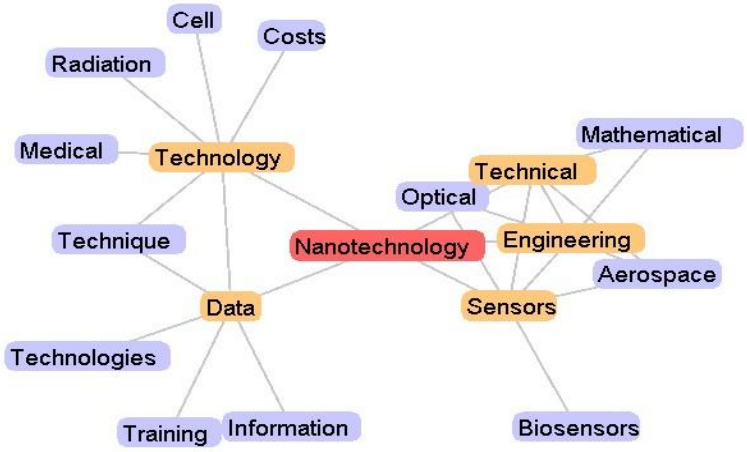


Fonte: Sérgio (2013)

Nos dois mapas pode-se verificar que os termos *Transgenic*, *Molecules* e *Cell* promovem a conexão entre os termos *Biotechnology* e *Genetic Engineering*.

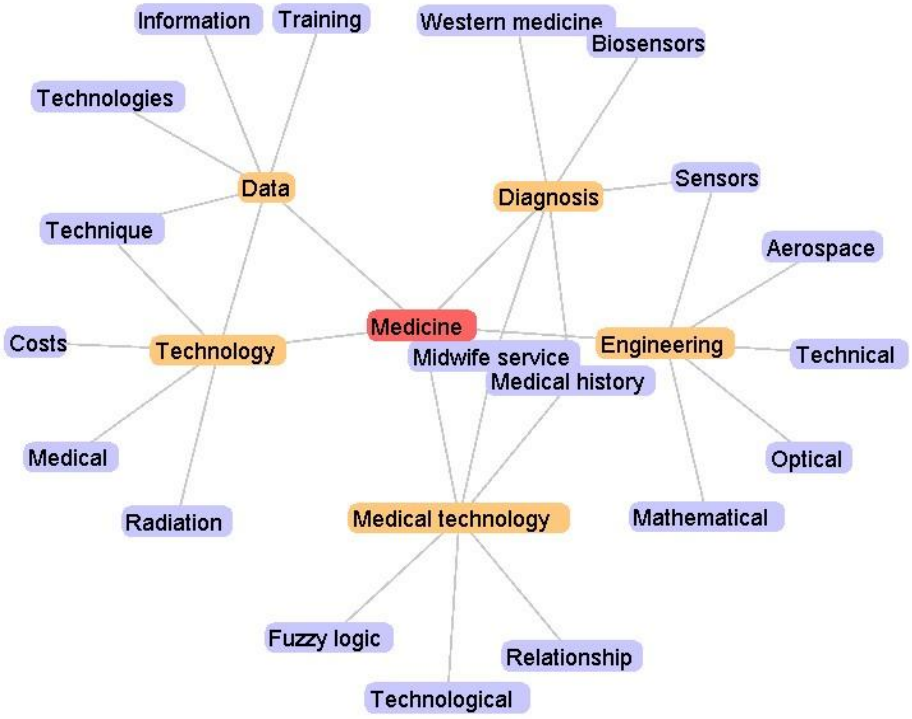
A Figura 7 e a Figura 8 apresentam cada uma mapas de tópicos obtidos a partir dos termos *Nanotechnology* e *Medicine*. Como é possível observar, os termos *Engineering*, *Technology* e *Data* promovem a conexão entre *Nanotechnology e Medicine*.

Figura 7 – Mapa de tópicos referente ao termo *Nanotechnology*



Fonte: Sérgio (2013)

Figura 8 – Mapa de tópicos referente ao termo *Medicine*



Fonte: Sérgio (2013)

A base de dados gerada para a condução dos estudos possibilita, ao nível de mapas de tópicos, análises temporais como as apresentadas no Gráficos 1 e no Gráfico 2 que forneçam uma visão estrutural de determinado domínio em função do tempo. Mapas de tópicos temporais podem promover indícios importantes sobre a evolução ou a retração de determinado domínio do

conhecimento visando à condução, por exemplo, de investimentos em pesquisa, desenvolvimento e inovação.

6 Considerações finais

A fim de desenvolver um modelo que possibilitasse a aplicação de técnicas de correlação e associação que considerasse a dimensão temporal e permitisse lidar com grandes volumes de dados não estruturados, o emprego da computação distribuída se mostrou essencial, ao passo que também demonstrou flexibilidade e escalabilidade. Seguindo por esta linha, buscou-se então a descoberta de relacionamentos entre termos que descrevessem determinado domínio de aplicação, e que fossem de caráter indireto e temporal.

Os resultados apresentados foram responsáveis por estabelecer a base para a geração das análises sobre o domínio pesquisado, sendo as mesmas explanadas por meio de gráficos temporais que tornam evidente a existência de padrões comportamentais entre os termos em questão. E com o intuito de melhorar o entendimento do domínio, foram gerados mapas de tópicos a partir dos vetores de contexto envolvendo determinado termo de análise.

Devido à aplicação do protótipo em um conjunto restrito de tempos e documentos – uma vez que não foi possível acessar completamente a base de dados *Science Direct*® – percebe-se a construção do cenário como uma das limitações existentes. Em função disso, a base de dados foi desenvolvida por meio de um especialista de domínio.

A partir da aplicação do processo de Descoberta de Conhecimento, padrões e tendências podem ser evidenciados. No âmbito de trabalhos futuros, vislumbra-se a evolução do software desenvolvido, gerando novas informações no processo e novas formas de visualização da informação, visando a descoberta de conhecimento por meio destes. Quanto a possíveis análises, destacam-se cenários onde haja a necessidade e a demanda por parte de organizações para compreender a competitividade ou a identificação de tendência de mercados, com dados oriundos da Web. Desta maneira, o conhecimento obtido pode

auxiliar no desenvolvimento e crescimento destas organizações, com o intuito de gerar vantagens competitivas e posicionamentos estratégicos.

Referências

AHMED, K.; MOORE, G. An introduction to topic maps. **The Architecture Journal**, [S.l.], v. 5, n. 5, jul. 2005. Disponível em: <http://msdn.microsoft.com/en-us/library/aa480048.aspx>. Acesso em: 31 ago. 2015.

BARALIS, E. et al. GraphSum: discovering correlations among multiple terms for graph-based summarization. **Information Sciences**, New York, v. 249, p. 96-109, nov. 2013.

CAO, L. et al. Flexible Frameworks for Actionable Knowledge Discovery. **IEEE Transactions on Knowledge and Data Engineering**, [S.l.], v. 22, n. 9, p. 1299-1312, set. 2010.

CHEN, Y. et al. Keyword search on structured and semi-structured data. In: ACM SIGMOD INTERNATIONAL CONFERENCE ON MANAGEMENT OF DATA, 1., 2009, Providence. **Proceedings**… Providence: ACM, 2009.

CHURCH, K. W.; GALE, W. A. Concordances for Parallel Text. In: ANNUAL CONFERENCE OF THE UW CENTRE FOR THE NEW OED AND TEXT RESEARCH, 8. 1991, Oxford. **Proceedings**… Oxford: [s.n], 1991. p. 40-62.

CHURCH, K. W.; HANKS, P. Word association norms, mutual information, and lexicography. **Computational Linguistics**, Cambridge, v. 16, n. 1, p. 22-29, mar. 1990.

CHURCH, K. W.; MERCER, R. L. Introduction to the Special Issue on Computational Linguistics Using Large Corpora. **Computational Linguistics**, Cambridge, v. 19, n. 1, p. 1-24, mar. 1993.

COHEN, T.; SCHVANEVELDT, R. W. The trajectory of scientific discovery: concept co-occurrence and converging semantic distance. **Studies in Health Technology and Informatics**, Amsterdam, v. 160, p. 661-665, 2010. Disponível em: <http://ebooks.iospress.nl/publication/13521>. Acesso em: 22 abr. 2016.

COHEN, T.; SCHVANEVELDT, R.; WIDDOWS, D. Reflective Random Indexing and Indirect Inference: A Scalable Method for Discovery of Implicit Connections. **Journal of Biomedical Informatics**, San Diego, v. 43, n. 2, p. 240-256, abr. 2010**.**

CONRAD, J. G.; UTT, M. H. A system for discovering relationships by feature extraction from text databases. In: ANNUAL INTERNATIONAL ACM SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL, 17., 1994, Dublin. **Proceedings**… New York: Springer-Verlag New York, 1994. p. 260-270.

COULOURIS, G. et al. **Sistemas distribuídos**: conceitos e projeto. 5. ed. Porto Alegre: Bookman, 2013.

EGGHE, L.; MICHEL, C. Strong similarity measures for ordered sets of documents in information retrieval. **Information Processing and Management**, [S.l.], v. 38, n. 6, p. 823-848, nov. 2002.

FENG, Y. Towards knowledge discovery in Semantic era. In: INTERNATIONAL CONFERENCE ON FUZZY SYSTEMS AND KNOWLEDGE DISCOVERY, 7., 2010, Yantai. **Proceedings**… Yantai: IEEE, 2010. p. 2071-2075.

FLEUREN, W. W. M.; ALKEMA, W. Application of text mining in the biomedical domain. **Methods**, [S.l.], v. 74, p. 97-106, mar. 2015.

GANTZ, J.; REINSEL, D. **The digital universe decade – are you ready?** Framingham: Idc – Iview, 2010. Disponível em: <https://www.emc.com/collateral/analyst-reports/idc-digital-universe-are-you-ready.pdf>. Acesso em: 21 abr. 2016.

HASHIMI, H.; HAFEZ, A.; MATHKOUR, H. Selection criteria for text mining approaches. **Computers in Human Behavior**, Oxford, v. 51, p. 729-733, out. 2015.

HILBERT, M.; LÓPEZ, P. The World's Technological Capacity to Store, Communicate, and Compute Information. **Science Magazine**, Washington, v. 332, p. 60-65, 01 abr. 2011. Disponível em: <http://science.sciencemag.org/content/332/6025/60>. Acesso em: 21 abr. 2016.

IVANOV, Nikita; DMITRIY, Setrakyan. **Real Time Big Data Processing with GridGain**. 2012.

JONES, W. P.; FURNAS, G. W. Pictures of relevance: a geometric analysis of similarity measures. **Journal of the American Society for Information Science**, [S.l.], v. 38, n. 6, p. 420-442, nov. 1987.

LIM, E. H. Y.; LIU, J. N. K.; LEE, R. S. T. Knowledge Discovery from Text Learning for Ontology Modeling. In: INTERNATIONAL CONFERENCE ON FUZZY SYSTEMS AND KNOWLEDGE DISCOVERY, 6., 2009, Tianjin. **Proceedings**… Tianjin: IEEE, 2009. p. 227-231.

MAIA, L. C. G.; SOUZA, R. R. Uso de sintagmas nominais na classificação automática de documentos eletrônicos. **Perspectivas em Ciência da**

Informação, **Belo Horizonte,** v. 15, n. 1, p.154-172, jan./abr. 2010. Disponível em: <http://portaldeperiodicos.eci.ufmg.br/index.php/pci/article/view/875/717>. Acesso em: 22 abr. 2016.

MANNING, C. D.; SCHÜTZE, H. **Foundations of statistical natural language processing**. Cambridge: MIT, 1999.

NOACK, T.; SCHMITT, I. Monitoring mobile cyber-physical systems by means of a knowledge discovery cycle. In: INTERNATIONAL CONFERENCE ON RESEARCH CHALLENGES IN INFORMATION SCIENCE (RCIS), 7., 2013, Paris. **Proceedings**… Paris: IEEE, 2013. p. 1-12.

RAMOS, H. de S. C.; BRÄSCHER, M. Aplicação da descoberta de conhecimento em textos para apoio à construção de indicadores infométricos para a área de C&T. **Ciência da Informação**, Brasília, v. 38, n. 2, p. 56-68, maio/ago. 2009. Disponível em: <http://revista.ibict.br/ciinf/article/view/1245/1423>. Acesso em: 22 abr. 2016.

RÊGO, B. L. **Gestão e governança de dados:** promovendo dados como ativo de valor nas empresas. Rio de Janeiro: Brasport, 2013.

SABOL, V. et al. Visual Knowledge Discovery in Dynamic Enterprise Text Repositories. In: INTERNATIONAL CONFERENCE INFORMATION VISUALISATION, 13., 2009, Barcelona. **Proceedings**… Barcelona: IEEE, 2009. p. 361-368.

SÉRGIO, Marina Carradore. **Uma arquitetura de descoberta de conhecimento baseada na correlação e associação temporal de padrões textuais**. 2013. 125 f. (Graduação em Tecnologias da Informação e Comunicação) Universidade Federal de Santa Catarina, Araranguá, 2013. Disponível em: <https://repositorio.ufsc.br/bitstream/handle/123456789/105488/TCC-Marina-Final_Com_Ficha.pdf?sequence=1&isAllowed=y>. Acesso em: 22 abr. 2016.

SHABBIR, A. et al. Predictive Data Mining and pattern recognition in the medical sector: Implementation and experience. In: WORLD CONGRESS ON COMPUTER APPLICATIONS AND INFORMATION SYSTEMS (WCCAIS), 1., 2014, Hammamet. **Proceedings**… Hammamet: IEEE, 2014.

SILVA, E. R. G.; ROVER, A. J. O Processo de descoberta do conhecimento como suporte à análise criminal: minerando dados da Segurança Pública de Santa Catarina. In: INTERNATIONAL CONFERENCE ON INFORMATION SYSTEMS AND TECHNOLOGY MANAGEMENT, 8., 2011, São Paulo. **Anais…** São Paulo: FEA, 2011.

SMALHEISER, N. R. Literature-Based Discovery: Beyond the ABCs. **Journal of the American Society for Information Science and Technology**, [S.l.], v. 63, n. 2, p. 218-224, fev. 2012. Disponível em:

<http://onlinelibrary.wiley.com/doi/10.1002/asi.21599/epdf>. Acesso em: 22 abr. 2016.

SOMASUNDARAM, G.; SHRIVASTAVA, A. **armazenamento e gerenciamento de informações:** como armazenar, gerenciar e proteger informações digitais. Porto Alegre: Bookman, 2011. 476 p.

SWANSON, D. R. Fish oil, Raynaud's syndrome, and undiscovered public knowledge. **Perspectives in Biology and Medicine**, Baltimore, v. 30, n. 1, p. 7-18, jan. 1986.

TANENBAUM, A. S.; STEEN, M. Van. **Distributed Systems**: principles and paradigms. 2. ed. Upper Saddle River: Prentice Hall, 2008.

VASHISHTHA, J.; KUMAR, D.; RATNOO, S. Revisiting Interestingness Measures for Knowledge Discovery in Databases. In: INTERNATIONAL CONFERENCE ON ADVANCED COMPUTING & COMMUNICATION TECHNOLOGIES, 2., 2012, Rohtak. **Proceedings…** Rohtak: IEEE, 2012. p. 72–78.

WOSZEZENKI, C. R.; GONÇALVES, A. L. Mineração de textos biomédicos: uma revisão bibliométrica. **Perspectivas em Ciência da Informação**, Belo Horizonte, v. 18, n. 3, p. 24-44, jul./set. 2013. Disponível em: <http://portaldeperiodicos.eci.ufmg.br/index.php/pci/article/view/1733/1189>. Acesso em: 22 abr. 2016.

WU, X. et al. Data Mining with Big Data. **IEEE Transactions on Knowledge and Data Engineering**, [S.l.], v. 26, n. 1, p. 97-107, jan. 2014.

YETISGEN-YILDIZ, M.; PRATT, W. A new evaluation methodology for literature-based discovery systems. **Journal of Biomedical Informatics**, San Diego, v. 42, n. 4. p. 633-643, ago. 2009. Disponível em: <https://faculty.washington.edu/melihay/publications/JBI_2010.pdf>. Acesso em: 22 abr. 2016.

ZHU, F. et al. Biomedical text mining and its applications in cancer research. **Journal of biomedical informatics**, San Diego, v. 46, n. 2, p. 200-211, abr. 2013. Disponível em: <http://www.sciencedirect.com/science/article/pii/S1532046412001712>. Acesso em: 22 abr. 2016.

Knowledge Discovery from Unstructured Information through Correlation and Association Techniques

Abstract: Nowadays, technology's current status seeks means to support the exponential increase of information available around the Internet or in organizations, and regarding that most of said information comes in a textual form, this is a challenge to the areas of crawling, storage, retrieval and analysis of information. This article aims to provide a Knowledge Discovery model based on the temporal correlation and association from large document collections. The methodology set for this process involve descriptive and explorative researches using papers taken right from the Science Direct® database as a tool for data collection and analysis. Through this kind of information is possible to extract rules, patterns, trends, and networks, all of them being usufel to the process of making decisions within organizations in order to generate competitive advantage. Thus, the main contribution of this paper relies on the proposition of a model towards the understanding of temporal aspects, considering factual relationships (through correlations) or not (through associations) between terms in a domain.

Keywords: Knowledge discovery. Correlation. Association. Unstructured information. Temporality.

Recebido: 22/10/2015

Aceirto: 11/04/2016

[1] Por "termos" compreendem-se palavras simples e/ou compostas que podem ser nomeadas/classificadas (chamados de entidades) ou não.