Medidas de similaridade em documentos eletrônicos[1]

Luiz Cláudio Gomes Maia (Escola de Ciência da Informação, UFMG)
Renato Rocha Souza (Escola de Ciência da Informação, UFMG)

Resumo: Algoritmos e técnicas aplicáveis na Recuperação da Informação na forma eletrônica estão em evolução e representam uma grande fatia dos estudos recentes em Ciência da Informação em conjunto com outras áreas como a Ciência da Computação. A Web, através de sua estrutura não linear formada por hiperlinks, ampliou as possibilidades, anteriormente limitada ao texto, de resultados mais satisfatórios com o uso da análise de ligações. A bibliometria é um exemplo do uso da análise de ligações. Este artigo faz uma compilação de técnicas de Recuperação de Informação e medidas de similaridade em um conjunto de documentos eletrônicos.

Palavras-chave: Agrupamento automático de documentos. Algoritmos de treinamento. Categorização de documentos. Similaridade.

Applicable algorithms and techniques in the Information Retrieval in the electronic form are in evolution and represent a great slice of the recent studies in Information Science in set with other areas as the Computer science. The Web, through its not linear structure formed by hyperlinks, extended the possibilities, previously limited to the text, of more satisfactory results with the use of the analysis of links. The bibliometric is an example of the use of the analysis of links. The research presents measured experiments involving of similarity in a set of electronic documents.

Keywords: Bibliometric. Link analysis. Similarity of electronic documents.

[1] Comunicação oral apresentada ao GT-08 - Informação e Tecnologia.

A informação cada vez mais é registrada diretamente em meios digitais. Vivencia-se uma consolidação, não só da convergência digital, mas também da criação de conteúdos totalmente já digitalizados. Neste contexto, a publicação e criação de conteúdos tornam-se mais fáceis e, por conseqüente, informações irrelevantes, de baixa qualidade e mesmo de baixa confiabilidade fazem parte de um "lixo informacional" crescente e que preocupa toda a sociedade.

Um dos principais campos de estudo da Ciência da Informação compreende o tratamento e organização da informação de forma a possibilitar resultados de busca satisfatórios, atendendo a demanda do usuário, sem a interferência do "lixo informacional".

Estudos apontam que no ano de 2007 existiam aproximadamente 550 bilhões de documentos on-line, com aproximadamente 7,5 petabytes entre *websites* e base de dados on-line (JANSSENS, 2007). Para armazenar 7,5 petabytes, em uma pilha de páginas de papel, onde cada página conteria 2500 caracteres, sendo que um byte equivale a 1 caractere, teríamos uma pilha de 300.000 km (1 cm para 100 páginas) o que daria para alcançar a lua ou dar a volta na terra 7,5 vezes. Uma pessoa lendo uma página por minuto gastaria 5.7 bilhões de anos para ler tudo. (JANSSENS, 2007). Com todo este volume informacional algum tipo de Sistema de Recuperação de Informação - SRI deverá ser necessário para que uma pessoa recupere rapidamente a informação que deseja em tempo satisfatório.

Para melhorar os processos relativos aos SRI, diversos grupos de pesquisa investem em avaliações e experimentos, sendo que as medidas estão presentes em toda pesquisa ou avaliação quantitativa de um SRI. As medidas existentes utilizadas em documentos, informações diversas e nos SRIs são conhecidas como métricas informacionais. No tratamento e organização da informação as métricas informacionais são importantes, pois permitem uma melhor atuação dos SRI. Estudos que apresentam novas métricas informacionais e mesmo melhorias e adaptações destas métricas para a informação digital são necessários. Muitas das métricas da Ciência da Informação foram pensadas quando não era possível o tratamento digital da informação. A bibliometria, campo de estudo que engloba métricas informacionais na Ciência da Informação, desenvolveu seu corpo teórico na década de 70 e estas métricas se mantêm até hoje inalteradas. O volume de informação e mesmo a técnica utilizada anteriormente era limitada ao processamento humano. Os sistemas computacionais permitem o processamento mais rápido destas métricas e abrem novas possibilidades de classificação e recuperação da informação.

Um Sistema de Recuperação da Informação (SRI) deve analisar os documentos para saber os itens de seu acervo que são relevantes frente a uma consulta do usuário. O objetivo é atender de forma satisfatória ao usuário. Para isto, pesquisas envolvendo técnicas e algoritmos aplicáveis em SRI são constantes. Atualmente, com todo o aporte computacional disponível, programas de computador podem aproveitar de um processamento rápido para melhorar ainda mais a satisfação do usuário no uso destes sistemas.

Algoritmos envolvendo métricas informacionais aplicáveis em SRI estão em evolução e representam uma grande fatia dos estudos recentes em Ciência da Informação, em conjunto com outras áreas como a Ciência da Computação.

Este artigo realiza uma compilação de técnicas atuais de Recuperação da Informação, e propõe medidas para realização de agrupamento por similaridade (*clustering*) e classificação de documentos eletrônicos.

Estas medidas permitem uma análise automatizada da similaridade de documentos eletrônicos, o que pode redundar em projetos inovadores de sistemas de recuperação de informações.

Para que um sistema de recuperação de informação possa responder às demandas dos usuários com tempos de respostas aceitáveis, é preciso que os documentos constantes da base

de dados sejam submetidos a um tratamento prévio. Este procedimento permite a extração dos descritores e sua estruturação com vistas a um acesso rápido às informações.

O metadado corresponde a um conjunto de informações que descreve o objeto digital. Sua correta criação e utilização é de extrema importância para a qualidade da informação e para os sistemas de recuperação de informação. Em relação à Web, na qual estão disponibilizados os periódicos eletrônicos, um dos problemas é que certos autores inflam os metadados de seus documentos para se ter uma visibilidade maior. Além disto, na maioria das páginas *html* quase não apresentam metadados, ou se constituem apenas das palavras chaves e descrição. (IRVIN, 2003, p.4). Tem-se trabalhado em padrões para realizar a correta formulação dos metadados, como por exemplo, o ETD-MS e o *Dublin Core* utilizados pela NDLTD. O *Dublin Core* também foi adotado como padrão pelo *Open Archives Initiative* e *D-Space* (GREENBERG, 2004). Porém quando se amplia do universo das bibliotecas digitais para toda a Web, se obtém que apenas 0,3% das páginas *html* contém metadados no padrão Dublin Core (LAWRENCE; GILES, 1999).

Ferramentas de extração e geração automática de metadados têm sido criadas e aprimoradas, podendo se citar como exemplo a *DC-Dot* e a *Klarity*, que criam metadados no padrão *Dublin Core* (IRVIN, 2003, p. 9; GREENBERG, 2004). Isto tem gerado uma discussão sobre qual seria o metadado com uma qualidade melhor: um gerado através da análise manual (humana) ou através da análise usando uma destas ferramentas (computador) (ANDERSON; PEREZ-CARBALLO, 2001).

O processo de indexação produzindo uma lista de descritores visa à representação dos conteúdos dos documentos. Ou seja, este processo tem como objetivo extrair as informações contidas nos documentos, organizando-as para permitir a recuperação destes últimos. Assim, os descritores deveriam ser obrigatoriamente, portadores de informação de maneira a relacionar um objeto da realidade extralingüística com o documento que traz informações sobre este objeto. Contudo, na maioria dos sistemas de recuperação de informação convencionais os descritores não passam de uma simples lista de palavras extraídas dos documentos que constituem as bases de dados.

Os itens de *pertinência* e *relevância* estão relacionados com a precisão e revocação dos sistemas de recuperação de informação, que por sua vez está relacionado com a consulta do usuário. Aumentando a precisão e/ou a revocação, obtemos o que qualquer usuário deseja: aumentar a possibilidade dele encontrar o que lhe seja relevante. Pensando ao extremo, o mundo perfeito para o usuário seria a criação de uma forma com que tudo o que fosse relevante para ele, ou seja, tudo o que a sua estratégia de busca representasse, fosse recuperada.

Necessita-se de uma solução atual, rápida e que seja aplicável a Web em sua forma atual. O modelo para similaridade, além de se aplicar a este contexto, deve também ser flexível para que possa se estender aos periódicos eletrônicos e às bibliotecas digitais de TDEs. Pretende-se alcançar esta solução aproveitando características de cada um destes conceitos de análise de ligações, como o algoritmo de PageRank e de estudos sobre Análise de Citação. Com estes dois últimos pretende-se medir os itens de *significância* e *atualidade*.

Como Salton (1968) afirma, o principal de todo e qualquer SRI deve ser aumentar a precisão (*precision*) e diminuir a revocação (*recall*), sendo:

$$precisão = \frac{número\ de\ doc.pertinentes\ recuperados}{número\ de\ documentos\ recuperados}$$

$$revocação = \frac{número\ de\ doc.pertinentes\ recuperados}{número\ total\ de\ documentos\ pertinentes}$$

Aumentando a precisão e/ou diminuindo a revocação, obtemos o que qualquer usuário deseja: aumentar a possibilidade dele encontrar o que lhe seja relevante. Talvez isso seja uma

utopia, mas o método proposto por essa análise, utilizando sintagmas nominais, pretende chegar o mais próximo dessa realidade o possível.

Análise de texto

A análise de texto (*text analysis*) corresponde a uma área que envolve outras subáreas como, por exemplo, a mineração de texto (*text mining)* e a área de Processamento de Linguagem Natural (PLN). A PLN também é uma subárea da inteligência artificial e da lingüística que estuda os problemas da geração e tratamento automático de línguas humanas naturais.

A Mineração de texto (*Text Mining*) refere-se ao processo de obtenção de informação a partir de texto em línguas naturais. Se praticada em conjunto com a mineração de dados, que consiste em extrair informação de bancos de dados estruturados; a mineração de texto extrai informação de dados não estruturados ou semi-estruturados.

O texto corresponde à principal parte das muitas que podem compor um documento, e seu tratamento, como um processo de criação dos índices, é explorado pelos SRIs.

Construção e armazenamento do Índice

O índice tem como objetivo a recuperação rápida da informação. A forma como se constrói, armazena e manipula o índice muda de acordo com a tecnologia empregada e conseqüentemente sua evolução. Tradicionalmente, CPUs eram lentas e a utilização de técnicas de compactação não seria interessante. Hoje as CPUs já são mais rápidas, entretanto temos um armazenamento em disco rígido lento, que para contornar necessitamos diminuir o espaço de armazenamento ou mesmo utilizar memórias mais rápidas (na hierarquia) como a RAM.

Basicamente a criação do índice significa criar um dicionário de palavras utilizadas em todos os documentos da coleção e criar um índice invertido indicando em qual documento cada palavra aparece.

Com a criação deste índice torna-se extremamente mais rápido a busca de informações do que recorrer a varrer todos os textos palavra por palavra.

A maior parte dos SRI tem como base o modelo clássico ou o modelo estruturado:

Nos modelos clássicos, cada documento é descrito por um conjunto de palavras-chave representativas, também chamadas de termos de indexação, que buscam representar o assunto do documento e sumarizar seu conteúdo de forma significativa. (BAEZA-YATES; RIBEIRO-NETO, 1999).

Nos modelos estruturados, podem-se especificar, além das palavras-chave, algumas informações acerca da estrutura do texto. Estas informações podem ser as seções a serem pesquisadas, fontes de letras, proximidade das palavras, entre outras.

Dentre os modelos clássicos, temos o booleano, o vetorial e o probabilístico. O modelo booleano é baseado na teoria dos conjuntos e possui consultas especificadas com termos e expressões booleanas. Nas consultas são utilizados operadores lógicos como E, OU, NÃO para filtragem do resultado.

Apesar de ser um modelo bastante simples e muito utilizado ele apresenta as seguintes desvantagens: (BAEZA-YATES; RIBEIRO-NETO, 1999)

- A recuperação é baseada numa decisão binária sem noção de casamento (*matching*) parcial;
- Nenhuma ordenação de documentos é fornecida;
- A passagem da necessidade de informação do usuário à expressão booleana é considerada complicada;
- As consultas booleanas formuladas pelos usuários são freqüentemente simplistas;

- Em conseqüência o modelo booleano retorna poucos ou muitos documentos em resposta às consultas;
- O uso de pesos binários é limitante;

Para contornar estas limitações novos modelos são desenvolvidos tendo como base algum destes modelos clássicos.

O modelo que permite localizar similaridade entre documentos é o vetorial. O vetor é definido através do conjunto de documentos que formam o corpora.

Todo o texto dos documentos são extraídos e convertidos em um formato que permita a fácil manipulação. Toda ordem das palavras é ignorada, o que pode ser interpretado como colocar todas as palavras de cada documento em um saco separado (a expressão *bag of words*). Todas as palavras em cada saco são contadas (processo de indexação) e o número de vezes que cada palavra aparece (forma mais simplista de dar valor ao peso) é armazenado em um vetor termo-por-documento.

Ele é arranjado de forma que cada linha representa uma palavra (termo) e cada coluna representa um documento. Os valores contem o peso dos termos para cada documento. Em geral este tipo de vetor e extenso e a maioria dos pesos dos termos são zero.

Tabela 1 – Exemplo do Modelo Vetorial

	d1	*d2*	*d3*	*d4*	*d5*	*d6*	*d7*	*D8*	*d9*
Rede	0	0,60	0	0,20	0,75	0,02	0	0,15	0,80
Social	0,20	0	0,05	0,30	0,75	0	0,02	0	0
Pesquisa	0	0,40	0	0,50	0	0	0	0	0,20
Vetor	0,20	0	0	0	0	0	0,10	0,10	0

Nas colunas estão representados os pesos de cada termo no documento. No exemplo acima o termo Rede tem o peso de 0,75 no documento 5 enquanto que o termo “Pesquisa” não aparece no documento 3 portanto seu peso é 0.

Sobre o uso de pesos no modelo vetorial, Baeza-Yates e Ribeiro-Neto (1999) apresenta algumas considerações:

- Pesos não binários podem considerar mais adequadamente *matchings* parciais;
- Estes pesos são utilizados para calcular um grau de similaridade entre a consulta e o documento;
- A fórmula com que são calculados os pesos varia dentre as implementações;

Cada documento (coluna) pode ser considerado como um vetor ou uma coordenada em um espaço do vetor do multidimensional em que cada dimensão representa um termo.

A medida *term frequency–inverse document frequency* (TF-IDF) corresponde a uma medida estatística utilizada para avaliar o quanto uma palavra é importante para um documento em relação a uma coleção (corpus). Esta importância aumenta proporcionalmente com o numero que de vezes que a palavra aparece no documento e diminui de acordo com o a freqüência da palavra na coleção.

O *term frequency* (TF) corresponde ao número de vezes que o termo aparece no documento. O que ocorre é uma normalização para evitar que documentos grandes se sobressaiam entre documentos pequenos no conjunto. A equação é dada por:

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

Onde:

Ni,j é o numero de ocorrências do termo no documento DJ e o denominador corresponde ao número de ocorrências de todos os termos no documento DJ.

Já o IDF é uma medida de grande importância para complementar a equação acima já que avalia a importância do termo na coleção. É obtida dividindo a quantidade de documentos pelo numero de documentos contendo o termo e então obtendo o logaritmo do resultado.

$$idf_i = log \frac{|D|}{|\{d_j : t_i \in d_j\}|}$$

Onde:

|D | é o total de documentos no corpus

$|\{d_j : t_i \in d_j\}|$ número de documentos onde o termo ti aparece.

Através da união das duas tem-se TF-IDF:

$$tfidf_{i,j} = tf_{i,j} . idf_i$$

Dependendo da aplicação e experimento, a partir do modelo TF-IDF podem surgir outros modelos que modificam a sistemática de atribuição de pesos.

A análise semântica Latente é uma técnica da PLN, relacionada a manipulação de vetores de índice. Ela esta relacionada a aplicação da matemática para analisar a relação entre termos e documentos e decompor o vetor de índice. O processo matemático utilizado é o SVD (*Simple Value Decomposition*).

Alguns autores e pesquisas também a chamam de *Latent semantic indexing* (LSI).

A LSA trabalha com a sinonímia e polissemia. Por exemplo, para a consulta "extravio de bagagem" feita a uma ferramenta de busca que usa LSA o sistema retornará documentos que contenham as frases "extravio de bagagem" e "extravio de mala" já que “bagagem” e “mala” têm o mesmo significado no contexto. Da mesma forma, em uma consulta por "banco de dados", o resultado da consulta incluirá somente documentos que contenham uma relação com "banco de dados", excluindo documentos que se referem à banco como objeto de descanso e banco como entidade financeira.

O LSA trabalha com vários vetores, criando desta forma uma matriz, que nas linhas estão representados os termos indexados de cada documento e nas colunas o documento, desta forma é criada a relação à matriz termo-documento. Explicando melhor esta relação, seja *ti* a linha e *dj* a coluna da matriz, e seja o elemento da matriz *Oij* que representaria o número de vezes que o termo i aparece no documento j.

Após de ser criada esta matriz termo-documento, é aplicado o SVD (*Simple Value Decomposition*), esta decomposição divide a matriz termo-documento em três matrizes: a matriz U que contém os termos, a matriz S que contêm os valores mais representativos da matriz termo-documento (os valores singulares) e a matriz V que contém os documentos. Depois de criadas estas três matrizes é escolhido um tamanho (nível k) para trabalhar com as mesmas. Escolhido este valor, são criadas três matrizes (que serão chamadas U', S' e V') de nível k, a estas três novas matrizes é multiplicado o vetor Q, que representa uma consulta. O resultado desta multiplicação será um vetor cujo conteúdo é uma lista dos documentos mais relevantes para a consulta fornecida.

Extração de descritores

De acordo RAMSDEN (1974, pág. 3) o termo linguagens naturais é comumente utilizado para denominar a linguagem falada e a linguagem escrita. É possível em indexação empregar a linguagem natural simplesmente como é falada ou usada nos documentos sem tentar, por exemplo, controlar sinônimos ou indicar os relacionamentos entre os termos. Um índice

feito desta maneira chama-se índice de linguagem natural. Como alternativa ao índice de linguagem natural pode-se usar uma linguagem artificial às necessidades do sistema de classificação, ou seja, uma linguagem de indexação. "*Esta linguagem refletirá um vocabulário controlado para o qual foram tomadas decisões cuidadosas sobre os termos a serem usados, o significado de cada um e os relacionamentos que apresentam*." (RAMSDEN, 1974, pág. 3)

Existem contextos na qual se pode utilizar uma linguagem de indexação: sistemas de classificação, listas de cabeçalhos de assunto, tesauros, etc. Sendo que elas consistem de um vocabulário controlado e uma sintaxe a ser seguida.

O processo de indexação visa à representação dos conteúdos dos documentos, produzindo uma lista de descritores. Ou seja, este processo tem como objetivo extrair as informações contidas nos documentos, organizando-as para permitir a recuperação destes últimos. Assim, os descritores devem ser, na maior extensão possível, portadores de informação, de maneira a relacionar um objeto da realidade extralingüística com o documento que traz informações sobre este objeto. Contudo, na maioria dos SRI convencionais, os descritores representam com muita limitação as informações presentes no documento.

Os termos isolados consistem na análise do vocabulário e da sintaxe utilizada dos itens a serem classificados e retirada e agrupamento dos termos que apresentam uma unidade semântica.

Alguns termos que podem prejudicar a recuperação, conhecidos como *stopwords*[2], são extraídos do texto através de um processo de tratamento do documento conforme ilustrado na figura abaixo.

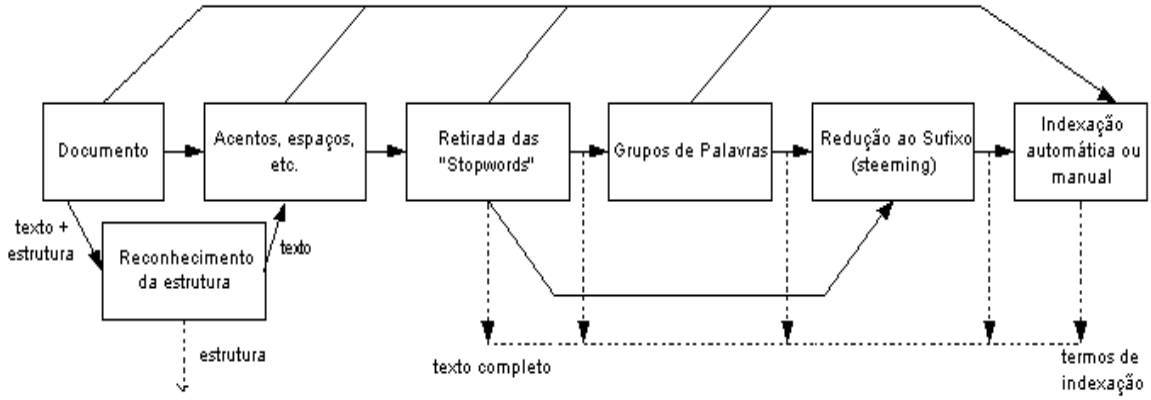

Figura 1 – Fases do processamento do documento para submissão a indexação.
Fonte: BAEZA-YATES & RIBEIRO-NETO, 1999, p. 166.

Ao final do processamento têm-se, através de um processo de indexação automática ou manual os termos de maior relevância para indexação. Técnicas como a de *Stemming*[3] podem ser utilizadas para reduzir a redundância semântica entre os termos.

A utilização das palavras como representação da descrição temática de um documento, segundo Kuramoto (2002), não é o ideal, devido aos vários problemas encontrados nas propriedades lingüísticas das mesmas. Exemplificando, temos:

---

[2] Palavras que não são úteis para recuperação de informações (e.g. palavras comuns, preposição, artigos, etc..)

[3] Processo de remover prefixos e sufixos das palavras do documento.

a) Polissemia: a palavra pode ter vários significados. Exemplo: chave (solução de um problema; ferramenta para abertura de portas; e também ferramenta para apertar parafusos);
b) Sinonímia: duas palavras podem designar o mesmo significado. Exemplo: abóbora e jerimum;
c) Duas ou mais palavras podem combinar-se em ordem diferente designando idéias completamente diversas. Exemplo: crimes, juvenis, vítimas (vítimas de crimes juvenis; vítimas juvenis de crimes).

A partir dessas três propriedades, Kuramoto conclui que a polissemia e a combinação de palavras podem atuar no resultado de uma busca em um SRI aumentando a taxa de ruído. No caso de ocorrência de sinonímia, pode ocorrer o incremento da taxa de silêncio. A taxa de Ruído e a Taxa de Silêncio correspondem a uma negação da taxa de Precisão e taxa Revocação já apresentadas. Temos:

$$taxa\ de\ ruído = \frac{número\ de\ doc.não\ pertinentes\ recuperados}{número\ total\ de\ documentos}$$

$$taxa\ de\ silêncio = \frac{número\ de\ documentos\ pertinentes\ não\ recuperados}{número\ total\ de\ documentos\ pertinentes}$$

Sintagmas Nominais e Verbais

KOCH e SILVA (1986) definem que o Sintagma Nominal consiste em um:

> "*conjunto de elementos que constituem uma unidade significativa dentro da oração e que mantêm entre si relações de dependência e de ordem. Organizam-se em torno de um elemento fundamental, denominado núcleo, que pode, por si só, constituir o sintagma*."

Na definição de KURAMOTO (1996), Sintagma Nominal "*é a menor parte do discurso portadora de informação*". Por exemplo, no termo "livro de bolso" a palavra "livro" constitui o centro do SN, ou o núcleo do SN; e "de bolso" caracteriza a identificação de uma classe de livros.

Alguns algoritmos de Recuperação da Informação trabalham com a identificação e indexação por lexemas para facilitar a recuperação de termos com sentido semelhante. Lexemas são palavras vinculadas através de uma relação denominada de flexão (PERINE, 1996). Por exemplo, as palavras cantora e cantar não compõem um lexema porque não pertencem à mesma classe morfológica. A primeira é substantivo e a segunda, verbo.

Já as palavras cantoria, cantilena e cantata não pertencem ao mesmo lexema porque embora apresentem um radical comum e pertençam à classe dos substantivos, diferem entre si por sufixos derivativos e não por sufixos flexivos.

As palavras cantor, cantora, cantores e cantoras compõem um lexema porque pertencem à mesma classe morfológica, a dos substantivos, e diferem entre si unicamente por sufixos flexivos (morfema zero, -a, -es, -as). Além disso, se distribuem de forma complementar.

Em classes de palavra que não variam, como as das preposições e conjunções, temos lexemas formados de uma única palavra.

Ao contrário das palavras, os SN não são símbolos sem referências, não restando dúvida que os SN possuem uma estrutura sintática, e que são portadores de uma estrutura lógica-semântica.

Para exemplificar, utilizaremos o sintagma nominal "O estudo da economia da informação". Este SN possui dois outros SN embutidos:

1. A economia da informação
2. A informação

Como se pode perceber, a descoberta dos SN evidencia uma organização em um esquema de árvore e, assim, diferentemente das palavras, o SN quando extraído de texto mantém o seu significado, o seu conceito.

A utilização dos sintagmas nominais como estrutura de acesso à informação contida em uma base de dados textual se apresenta como uma alternativa aos sistemas tradicionais de recuperação de informação. Podendo aproximar um pouco mais a necessidade informacional do usuário da forma como os documentos potenciais (aqueles que talvez possam responder a essa demanda) possam estar representando essa necessidade.

O Sintagma Verbal é caracterizado pela presença do verbo. Além do verbo, outros termos podem fazer parte do SV, dependendo do verbo que funciona como núcleo. Esses outros elementos são, por sua vez, sintagmas - nominais ou preposicionados.

A Entretanto uma classificação e recuperação da informação somente o uso de lexemas não é interessante (PERINE, 1996). Para PERINE o uso de sintagma nominal é mais eficiente neste processo.

A metodologia proposta por SOUZA (2005) apresenta os seguintes passos:

- Extração dos Sintagmas Nominais do texto.
- Analise de cada um dos sintagmas nominais e cálculo da pontuação de cada um como descritor.

Para esta avaliação, relacionam-se a relevância dos SN como descritores e os fatores:
a) freqüência de ocorrência do SNs no texto do documento.
b) a incidência dos SN no conjunto de documentos.
c) seus níveis.
d) suas estruturas sintáticas.
e) sua ocorrência no tesouro da área.
A pontuação seguirá a seguinte tabela:

Tabela 2 – Pontuação da avaliação dos SNs como descritores

1	SN extremamente relevante como descritor (SNER)
0,5	SN razoavelmente relevante como descritor (SNRR)
0,25	SN moderadamente relevante como descritor (SNMR)
0 S	SN não relevante como descritor

Fonte: Souza, 2005.

Pontuação da avaliação da qualidade dos descritores de um documento:

$$pontuação\ (descritor) = (num.SNER) + (0{,}5(num.SNRR)) + (0{,}25(SNMR))$$

Classificação

A classificação está presente em todas as nossas atividades do cotidiano. Isto é, de certa forma, comprovado na neurociência onde já se estabelece o processo de associação ou associar como o processo básico de funcionamento do cérebro humano.

> *"só ela nos permite orientar-nos no mundo a nossa volta, estabelecer hábitos, semelhanças e diferenças, reconhecer os lugares, os espaços, os seres, os acontecimentos; ordená-los, agrupá-los, aproximá-los uns dos outros, mantê-los em conjunto ou afastá-los irremediavelmente." (POMBO, 2003, p.01)*

O processo básico de classificação pode ser resumido em associar dois itens. O processo final de classificação corresponde a uma atividade anterior de associação. Lakoff (1987)

afirma que “sem a capacidade de categorizar, nós não poderíamos atuar nem no mundo físico nem no nosso mundo social e intelectual”.

Processos classificatórios existem desde a antiguidade onde organizar o conhecimento humano era preocupação dos filósofos, entretanto a palavra classificar, que vem do latim *classis,* teve sua origem em 1733 combinando *classis* e *facere,* e somente no final do século XVII foi empregada para referir-se a ordem da ciência e do conhecimento.

> *“...num sentido geral, é reunir em classes ou grupos, que apresentam entre si certos traços de semelhança, ou mesmo de diferença. Podemos ainda dizer que a classificação é um processo mental por meio do qual podemos distinguir coisas, pelas suas semelhanças ou diferenças, estabelecer as suas relações e agrupá-las em classes de acordo com essas relações.” (SOUZA, 1950, p.3)*

A ciência há tempos é utilizada como base para elaboração da classificação do conhecimento. Por tratar na maioria das vezes de forma empírica com seu processo de desenvolvimento a ciência e a comunidade dependem da classificação como inclusive a própria comunicação cientifica. Conforme Kwasnik:

> “*O processo da descoberta e a criação do conhecimento na ciência seguiu tradicional ao trajeto da exploração, observação, descrição, análise, e síntese sistemática e testar dos fenômenos e dos fatos, conduzidas toda dentro da estrutura de uma comunicação de uma comunidade de pesquisa particular com seus metodologia e conjunto aceitados das técnicas*” (*KWASNIK, 1999, pág. 22)*

Atualmente vivemos em uma época em que os modelos de classificação tradicionais são criticados por incentivarem uma ciência atomizada onde os conhecimentos gerados não se convergem.

A Ciência da Informação adota em sua classificação bibliográfica (a classificação utilizada para organização de uma coleção de livros) muitos conceitos herdados da classificação do conhecimento. A classificação bibliotecária como disciplina acadêmica tem apenas 125 anos e seu ensino e pesquisa têm crescido lentamente. (SATIJA, 2000, pág 221)

A classificação aplicada à prática da biblioteconomia corresponde a dar aos livros e documentos, de um modo geral, o lugar certo em um sistema de recuperação de informações, onde existe uma coleção que abrange os vários campos do saber sendo cada item agrupado ou representado conforme sua semelhança, diferenças e relações recíprocas com outros itens dentro da coleção. A classificação pode corresponder ainda em determinar o assunto de um documento. Ou também, traduzir os assuntos dos documentos da linguagem natural para a linguagem artificial, de indexação, de forma a ser utilizada num sistema que permita recuperar eficientemente informações.

Aprimoramentos sobre a classificação automática, ou seja, realizada sem a intervenção do homem tornam-se cada vez mais importantes num mundo com crescimento exponencial do volume de informação.

Nesta pesquisa utilizam-se algoritmos e medidas de similaridades para realizar uma classificação automatizada de documentos eletrônicos. Por ser a classificação uma atividade inerente ao ser humano técnicas e algoritmos computacionais tentam se aprimorar obtendo resultados próximos de uma classificação feita pelo homem, nesta busca algumas técnicas inclui até o uso de inteligência artificial.

É importante definir o tamanho da estrutura do sistema de classificação (ou seja o número de classes) de acordo com o tamanho da coleção (SVENONIOUS, 1985, pág. 11). As classes principais da estrutura são de extrema importância para a boa organização e uso do sistema classificatório.

A classificação automática toma como base as propriedades do objeto que se pretende classificar e através delas define a(s) classes(s) a qual pertence. Ao classificar que um docu-

mento é similar a outro é necessário realizar um processo de associação entre estes documentos. Um documento com metadados (incluindo descritores) torna o processo de classificação automática mais eficaz. (SVENONIOUS, 1985, pág. 13)

Clustering corresponde as técnicas que permitem subdividir um conjunto de objetos em grupos. O objetivo é fazer que cada grupo (ou cluster) seja o mais homogêneo possível levando em consideração que os objetos do grupo tenham propriedades similares e que os objetos nos outros grupos sejam diferentes (JANSSENS, 2007).

O algoritmo de agrupamento (*clustering)* pode funcionar, basicamente, através de duas formas:

- Número de agrupamentos automático – o número de categorias é definido automaticamente, geralmente com base no número de documentos da coleção.
- O número de clusters é pré-definido e as categorias apresentadas – As categorias já se encontram definidas antes da execução do algoritmo. Esta definição pode ser dada a partir de um conjunto de treinamento (*training set*) que correspondem a documentos já classificados que servirá de base para o algoritmo classificar novos documentos.

No Cadê um dos primeiros portais de pesquisa brasileiros apresenta documentos organizados em diretórios, e a classificação de novos documentos bem como a criação de categorias é feita manualmente ou pelo próprio usuário na hora da inserção do documento (site) no diretório. Esta característica faz com que a definição de um diretório apenas pelo seu nome seja suficiente o que não ocorre no caso de uma classificação automática. O computador necessita de mais informações além do nome para se basear a classificação.

O *Computer Science Research Paper Search Engine* (CORA)[4] (McCALLUM, 2000), é um projeto de um portal para pesquisa de artigos na área da Ciência da Computação que tem a organização de seus diretórios feita de forma manual, porém como a classificação de documentos é feita de forma automática, seus diretórios recebem uma definição, que é criada por meio de um conjunto de treinamento, algumas palavras-chave, atribuídas manualmente às categorias, e por um algoritmo de treinamento (o algoritmo *Bootstrapping*), que refina a definição das categorias.

Medidas de similaridade em documentos eletrônicos

Os algoritmos que retornam similaridade entre documentos trabalham com métricas que retornam o quanto um documento é similar a outro. Existem diversos algoritmos e métricas utilizados em fins diversos. Um algoritmo deste tipo pode, por exemplo, ser utilizado na grade de programação digital da televisão para fornecer programas similares ao gosto do usuário, conforme demonstrado por Fabio (SANTOSSILVA, 2005) em projeto denominado Sistema de Recomendação Personalizada de Programas de TV (SRPTV).

No campo da estatística temos duas medidas de similaridade básicas que se expande para outros estudos: correlação e coseno. A correlação de Person (ou só correlação) entre dois vetores retorna um valor entre 0 e 1. Se for 1 eles estão fortemente correlacionados, isto é, os valores de um vetor podem predizer os valores do outro. Se for 0 não existe correlação. E se for -1 existe uma correlação inversamente proporcional.

[4] O site do projeto não está atualizado e aparenta ter sido descontinuado.

O coseno é similar à correlação, retornando valores entre 0 e 1. Ele mede o ângulo entre dois vetores num espaço vetorial. Quanto mais próximo de 1 for o valor, mais similares são os dois vetores.

Para se localizar a similaridade entre dois documentos em um SRI utilizando VSM, calcula-se o coseno do ângulo formado no vetor termo-por-documento. No VSM padrão quanto menor o ângulo, mais próximo de 1 será o coseno e mais similar será o documento em relação a aquele termo.

$$sim(\vec{d_1}, \vec{d_2}) = \cos(\widehat{\vec{d_1}\vec{d_2}}) = \frac{\vec{d_1} \cdot \vec{d_2}}{|\vec{d_1}| \cdot |\vec{d_2}|} = \frac{\sum_i w_{i,1} \cdot w_{i,2}}{\sqrt{\sum_i w_{i,1}^2} \cdot \sqrt{\sum_i w_{i,2}^2}}$$

Onde: $w_{i,j}$ é o peso do temo t_i no documento d_j

Baeza-Yates e Ribeiro-Neto (1999) nos apresenta algumas outras observações sobre este modelo como um todo:

- Um conjunto ordenado de documentos é retornado, fornecendo uma melhor resposta à consulta.
- Documentos que têm mais termos em comum com a consulta tendem a ter maior similaridade;
- Aqueles termos com maiores pesos contribuem mais para o casamento do que os que têm menores pesos;
- Documentos maiores são favorecidos;
- A similaridade calculada não tem um limite superior definido.

O uso de um SRI e de um algoritmo de *clustering* para agrupar documentos envolve calcular a distância entre estes documentos na matriz. Existem além do co-seno de similaridade de outras medidas, sendo que a distancia Euclidiana é também muito utilizada. A distância Euclidiana entre dois documentos d1 e d2 é definida por:

$$d(\vec{d_1}, \vec{d_2}) = \sqrt{\sum_i (w_{i,1} - w_{i,2})^2}$$

Onde: $w_{i,j}$ é o peso do temo t_i no documento d_j.

A distância euclidiana necessita que quatro condições, nos vetores x, y e z, sejam validas para atuar como medida:

1. $d(x, y) \geq 0$
2. $d(x, x) = 0$
3. $d(x, y) = d(y, x)$
4. $d(x, z) \leq d(x, y) + d(y, z)$

Mais uma vez o tamanho do documento tem grande influência quando se utiliza a distância euclidiana.

O algoritmo de Rocchio (ROCCHIO, 1971 citado por HARMAN, 1992) é um algoritmo de lote, que produz um novo vetor de pesos w a partir de um vetor de pesos existente w_1 e de um conjunto de exemplos de treinamento. O j^{th} componente w_j do novo vetor de componentes é (LEWIS, 1996):

$$w_j = \alpha w_{1,j} + \beta \frac{\sum_{i \in C} x_{i,j}}{n_C} - \gamma \frac{\sum_{i \notin C} x_{i,j}}{n - n_C}$$

Sendo:

$$C = \{1 \leq i \leq n : y_i = 1\}$$

Onde n é o número de exemplos de treinamento, C é o conjunto de exemplos de treinamento positivos e é número de exemplos positivos de treinamento. Os parâmetros α, β e ◎ controlam o impacto relativo do vetor de pesos original, dos exemplos positivos e dos negativos respectivamente.

O algoritmo de Rocchio baseia-se na satisfação através do *feedback* do usuário com os resultados apresentados (treinamento positivo). Pode-se fazer uma relação com as técnicas de *Relevance Feedback* apresentadas e discutidas por Buckley (1995).

Para Buckley, *Relevance Feedback* é o processo automático de refinamento de uma consulta inicial, utilizando informações fornecidas pelo usuário sobre a relevância dos documentos previamente recuperados (em uma consulta anterior).

E é através deste processo de retro-alimentação, que corresponde a aplicar a equação apresentada, serão obtidas definições cada vez mais apuradas para as categorias envolvidas.

A medida kNN é definida por Yang em 1994 (apud CALADO et all, 2006) e definida por este nome na pesquisa de Calado (et all 2006) devido a se basear em testes realizados com categorias (k) vizinhas (*nearest neighbor*) e através de um processo de afunilamento definir a categoria.

A seguinte equação ilustra o algoritmo kNN:

$$S_{c_i,d} = \sum_{d' \in N_k(d)} similaridade\,(d,d') f(c_i,d')$$

Onde

K é igual ao número de vizinhos, *Nk*(*d*) corresponde aos documentos mais similiares a *k. e f*(*ci, d_*) corresponde a uma função binária que retorna se o documento *d'* pertence a uma categoria *ci* ou não.

O objetivo é filtrar os documentos baseado na predominância dos k vizinhos mais próximos. Os vizinhos mais próximos são os documentos que possuem maior valor de similaridade.

Algumas métricas utilizadas para identificação de dados similares são *Edge Cover*, *Shingsem*, *shingcom*, Distância de edição, *Similarity flooding*, *Shingles* e Serie temporal.

Muitos algoritmos de agrupamento requerem como parâmetro predefinido o número de grupos, ou então outro parâmetro para definir a granularidade. A definição do número de grupos pode apresentar dificuldades de acordo com o conjunto de medidas e técnicas utilizadas. Existem alguns métodos e algoritmos para definir a quantidade de grupos de forma automática. Como por exemplo: método baseado na distância, *dendrogram*, *Curvas de Sihouette*, *Bem-Hur*, *Elisseeff* e *Guyon*.

Pesquisas Similares

Esta pesquisa utiliza um modelo proposto por SOUZA (2005) onde propõe o uso de sintagmas nominais como descritores para recuperação de documentos.

Calado (et all, 2006) realiza um experimento utilizando as medidas de similaridade: *Amsler*, *Bibliographic Coupling*, Co-Citacion, kNN, SVM e *Naive Bayes* utilizando um corpora baseado no diretório de busca CADE. A pesquisa conclui que são necessárias novas experiências em outros corpos de documentos.

Referências

ANDERSON, J.,; PEREZ-CARBALLO, J.. The Nature of Indexing: How Humans and Machines Analyze Messages and Texts for Retrieval. Part I: Research, and the Nature of Human Indexing. Information Processing and Management, n. 37, 2001. p. 231-254.

BAEZA-YATES, R.; RIBEIRO-NETO, B.. Modern Information Retrieval. New York: ACM Press, 1999.

BUCKLEY, C.; SALTON, G.. Optimization of Relevance Feedback Weights In: Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Washington: USA. 1995. p. 9-13.

CALADO, P. P.; CRISTO, M.; MOURA, E. S.; GONÇALVES, M. A.; ZIVIANI, N.; RIBEIRO-NETO, B.. Linkage similarity measures for the classification of Web documents. Journal of the American Society for Information Science and Technology (JASIST), vol. 57, no. 2, 2005. p. 208-221.

GREENBERG, J.. Metadata Extraction and Harvesting: A Comparison of Two Automatic Metadata Generation Applications. Journal of Internet Cataloging, 6(4), p. 59-82, 2004.

HARMAN, D.. Relevance feedback and other query modification techniques. In William B. Frakes and Ricardo Baeza-Yates, editors, Information Retrieval: Data Structures and Algorithms, p. 241-263. Prentice Hall, 1992.

IRVIN, K.K.. Comparing Information Retrieval Effectiveness of Different Metadata. Generation Methods. A Master's paper for the M.S. in I.S. degree. April, 2003.

JANSSENS, F.. Clustering of scientific fields by integrating text Mining and bibliometrics, Katholieke Universiteit Leuven: Faculteit Ingenieurswetenschappen. Mei, 2007.

KOCH, I. V.; SILVA, M.C.P.S.. Lingüística aplicada ao português: sintaxe. São Paulo, Cortez, 1985.

KURAMOTO, H.. Sintagmas Nominais: uma nova proposta para a Recuperação da Informação. DataGramaZero, v. 3, n. 1, fev. 2002.

_____. Uma abordagem alternativa para o tratamento e a recuperação da informação textual: os sintagmas nominais. Ciência da Informação, Brasília, p. 182-192, v. 25, n. 2, maio/ago. 1996.

KWASNIK, B.H.. The role of classification in knowledge representation and Discovery. Library Trends, p. 22-47, v. 48, n. 1, Summer, 1999.

LAKOFF, G.. Women, fire, and dangerous things: what categories reveal about the mind. Chicago: The University of Chicago Press, 1987.

LAWRENCE, S.; GILES, C.. Accessibility of Information on the Web. Nature, p.107-109, n. 400, 1999.

MCCALLUM, A. K.; et al.. Automating de Construction of Internet Portals with Machine Learning Information Retrieval, p. 127-163 , v.2, n. 3, 2000.

PERINE M.A.; et al.. O Sintagma Nominal em Português: Estrutura, Significado e Função, Revista de Estudos da Linguagem. n. esp.. 1996.

POMBO, O.. Da Classificação dos Seres à Classificação dos Saberes, Leituras. Revista da Biblioteca Nacional de Lisboa, n. 2, Primavera, pp. 19-33. disponível no site: http://www.educ.fc.ul.pt/docentes/opombo/investigacao/opombo-classificacao.pdf consultado em 05/12/2003

RAMSDEN, M. J.. An introduction to index language construction, a programed text. London: C. Bingley, 1974.

SALTON, G.. Automatic information organization and retrieval. New York: McGraw-Hill, 1968.

SANTOS SILVA, F.. Personalização de Conteúdo na TVDI através de um Sistema de Recomendação Personalizada de Programas de TV (SRPTV). Anais... III Fórum de Oportunidades em Televisão Digital Interativa, Poços de Caldas, 2005.

SATIJA, M.P.. Library classification:an essay in terminology. Knowledge organization, p. 221-229, v. 27,n. 4, 2000.

SOUZA, J.S.. Classificação: sistemas de classificação bibliográfica. 2.ed. São Paulo: Departamento Municipal de Cultura, 1950.

SOUZA, R.R.. Uma proposta de metodologia para escolha automática de descritores utilizando sintagmas nominais. Tese de Doutorado. Orientadora Profª. Lidia Alvarenga. ECI: UFMG, 2006.

SVENONIOUS, E.. Classification theory. March, 1985. 19p