Uso de sintagmas nominais na classificação automática de documentos eletrônicos[1]

Luiz Cláudio Maia

Doutor e mestre em Ciência da Informação pela Universidade Federal de Minas Gerais. Professor e coordenador de cursos da Faculdade Pitagoras. Diretor da Alarmsoft Tecnologia em Segurança.

Renato Rocha Souza

Doutor em Ciência da Informação
Professor Adjunto da Escola de Ciência da Informação

Esta pesquisa verificou se ocorre aprimoramento na classificação de documentos eletrônicos com o uso de técnicas e algoritmos de mineração de texto (análise de texto) utilizando-se, além das palavras, sintagmas nominais como indexadores. Utilizaram-se duas ferramentas nos experimentos propostos desta pesquisa o OGMA e a WEKA. O OGMA foi desenvolvido pelos autores para automatizar a extração dos sintagmas nominas e o cálculo do peso de cada termo na indexação dos documentos para cada um dos seis métodos propostos. A WEKA foi utilizada para analisar os resultados encontrados pelo OGMA utilizando aos algoritmos de agrupamento e classificação, SimpleKMeans e NaiveBayes, respectivamente, obtendo um valor percentual indicando quantos documentos foram classificados corretamente. Os métodos com melhores resultados foram o de termos sem stopwords e o de sintagmas nominais classificados e pontuados como descritores.

Palavras-chave: *Análise de texto; Agrupamento automático de documentos; Indexação automática; Sintagmas nominas.*

[1] Agradecemos ao CNPq pelo apoio a esta pesquisa.

Use of noun phrases in automatic classification of electronic documents

This research work presents a proposal for the classification of electronic documents using techniques and algorithms based on natural language processing and noun phrases indexing along with plain keywords. Two tools, OGMA and Weka, were used for the experiments proposed. OGMA was developed by the author to automate the extraction of noun phrases and to perform the calculation of the weight of each term in the process of document indexing for each of the six proposed methods. The WEKA was used to analyze the OGMA results using the algorithms of clustering and classification "Simplekmeans" and "NaiveBayes", respectively. This process resulted in a percentage value indicating how many documents were classified correctly. The best performing methods were those with the terms without stopwords and the classified and scored noun phrases.

Keywords: *Text analysis; Clustering; Automatic indexing; Noun phrases; Natural language processing*

Recebido em 10.06.2009 Aceito em 05.**03.2010**

1 Introdução

Vivencia-se com o advento das redes de comunicação uma crescente preocupação com essas formas de tratamento e organização da informação, pois cresce de maneira cada vez mais rápida a quantidade dos textos armazenados em formato digital. A maioria deles é esquecida, pois nenhum ser humano pode ler, entender e sintetizar todo esse volume informacional.

Isso tem incentivado os pesquisadores a explorar estratégias para tornar acessível, ao usuário, a informação relevante (BAEZA-YATES; RIBEIRO-NETO, 1999; FRANTZ; SHAPIRO; VOISKUNSKII, 1997; MEADOW; BOYCE; KRAFT, 2000; CROFT, 2000).

Com todo esse volume informacional somente o tratamento do texto dos documentos já não constitui garantias de uma recuperação eficiente. O aprimoramento e também novos estudos envolvendo a análise pelo computador da linguagem humana são necessários para melhorar a recuperação desses documentos. A análise realizada pelo computador da linguagem humana é conhecida como processamento da linguagem natural (PLN).

Este artigo apresenta o resultado de um experimento realizado em uma tese (MAIA, 2008) na qual foi desenvolvida uma ferramenta computacional que adaptasse e complementasse os modelos propostos por três pesquisas importantes em ciência da informação, computação e lingüística (SOUZA, 2005; MIORELLI, 2001; PERINI, 1996). Essa ferramenta computacional foi denominada Ogma e trata-se de um programa que permitiu realizar todos os cálculos e análises textuais necessárias para a execução do experimento proposto.

2 Recuperação da Informação

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

O processo de indexação produzindo uma lista de descritores visa à representação dos conteúdos dos documentos. Ou seja, esse processo tem como objetivo extrair as informações contidas nos documentos, organizando-as para permitir a recuperação destes últimos. Contudo, na maioria dos sistemas convencionais de recuperação de informação, os descritores não passam de uma simples lista de palavras extraídas dos documentos, que constituem a coleção.

De acordo com Ramsden (1974), 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 dessa maneira chama-se índice de linguagem natural. Como alternativa ao índice de linguagem natural, pode-se usar uma linguagem artificial adaptada às necessidades do sistema de classificação, ou seja, uma linguagem de indexação. "Esta linguagem refletirá em 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, p. 3).

Existem contextos nos quais se pode utilizar uma linguagem de indexação: sistemas de classificação, listas de cabeçalhos de assunto, tesauros etc.; sendo que as linguagens 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, esse processo tem como objetivo extrair as informações contidas nos documentos, organizando-as para permitir a recuperação destes últimos. Assim, os descritores devem, na maior extensão possível, ser portadores de informação, de maneira a relacionar um objeto da realidade extralingüística com o documento que traz informações sobre esse objeto. Contudo, na maioria dos Sistemas de Recuperação de Informação (SRI)

convencionais, os descritores representam com muita limitação as informações presentes no documento.

Os termos isolados decorrem da análise do vocabulário e da sintaxe dos documentos a serem classificados e conseqüente extração e agrupamento dos termos que apresentam uma unidade semântica.

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

- polissemia: a palavra pode ter vários significados. Exemplo: chave (solução de um problema; ferramenta para abertura de portas; e também ferramenta para apertar parafusos);
- sinonímia: duas palavras podem designar o mesmo significado. Exemplo: abóbora e jerimum;
- 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 (2002) 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.

3 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 PLN, sendo a ultima uma subárea da inteligência artificial e da lingüística que estuda os problemas da geração e tratamento automático de línguas naturais.

A mineração de texto constitui na extração de informações sobre tendências ou padrões em grandes volumes de documentos textuais, em que uma amostra significativa de informações seja avaliada em textos contidos em bases textuais e em fontes de informação em linha. (POLANCO; FRANÇOIS, 2000)

A lingüística é o estudo científico da linguagem verbal humana. A gramática gerativa é uma teoria lingüística elaborada por Noam Chomsky e pelos lingüistas do *Massachusetts Institute of Technology* (*MIT*) entre 1960 e 1965. O estudo da gramática generativa revolucionou os estudas da lingüística.

Os modelos tradicionais descrevem somente as frases realizadas e, portanto, não relacionam um grande número de dados lingüísticos (como a ambigüidade, os constituintes descontínuos, etc.). Nessa perspectiva, esses modelos tradicionais correspondem a um mecanismo finito que permite analisar um conjunto de frases (bem formadas, corretas) de uma língua, e somente elas.

Chomsky (1969) propõe então uma teoria capaz de dar conta da criatividade do falante, de sua capacidade de emitir e de compreender frases inéditas.

A gramática gerativa compreende na realidade um conjunto de modelos teóricos que tem em comum a sua intenção de estudar o dispositivo mental inato, responsável pela produção lingüística.

Na gramática gerativa aplicam-se três modelos mais conhecidos: (CHOMSKY, 1969; LOPES, 1999):

1. modelo dos estados finitos;
2. modelo sintagmático – gramática sintagmática;
3. modelo transformacional – gramática transformacional (GT).

O modelo dos estados finitos baseia-se na aplicação de regras recursivas sobre um vocabulário finito. Segundo Chomsky, esse tipo de descrição gramática concebe as frases como tendo sido engendradas por uma série de escolhas executadas pelo formulador da frase.

Na mesma linha, segundo Trask (2004, p. 87), a gramática sintagmática é...

> [...] um tipo de gramática gerativa, que representa diretamente a estrutura dos constituintes. Normalmente, consideramos a estrutura de qualquer sentença (frase) como um caso de estrutura de constituintes, em que as unidades menores se combinam para formar unidades maiores, que são, por sua vez, combinadas formando unidades ainda maiores, e assim sucessivamente.

Assim, por exemplo, na gramática sintagmática, o sujeito da frase é identificado por propriedades formais. Nesse modelo toda a oração é formada por unidades de significado (os sintagmas) que se organizam de acordo com leis determinadas. As leis que organizam os sintagmas são chamadas sintagmáticas.

Conforme Perini (1996), o sintagma nominal possui uma estrutura bastante complexa, pois é possível distinguir em sua composição várias funções sintáticas. Seu núcleo pode ser um nome (comum ou próprio) ou um pronome (pessoal, demonstrativo, indefinido, interrogativo ou possessivo). O sintagma nominal pode também ser constituído por determinantes e/ou modificadores, sendo que os modificadores antecedem ou sucedem o núcleo, enquanto os determinantes apenas o antecedem (MIORELLI, 2001).

Alguns autores levam o conceito de sintagma nominal da lingüística para trabalhar as questões semântica e informacional presentes no sintagma nominal. Silva e Koch (2007, p. 18-19) 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.

Também na definição de Kuramoto (1996), sintagma nominal "é a menor parte do discurso portadora de informação".

Como exemplo, o sintagma nominal:

"O estudo da economia da informação".

Possui três outros SN embutidos:

1. a economia da informação
2. a informação
3. a economia

Os SN "a informação" e "a economia" são sintagmas nominais aninhados dentro do SN "a economia da informação".

A extração automatizada de sintagmas nominais sempre foi um elemento de dificuldade em pesquisas de recuperação da informação envolvendo sintagmas nominais. Kuramoto (2002) ressalta essa dificuldade, principalmente em um grande volume de textos:

> O processo de reconhecimento, extração e indexação não automatizada, além de ser inviável economicamente em se tratando de grandes volumes de documentos, pode prejudicar a uniformidade no processo de reconhecimento, extração e indexação dos sintagmas nominais (KURAMOTO, 2002).

O pesquisador complementa no mesmo artigo que "A inexistência dessas ferramentas impede uma avaliação mais consistente envolvendo amostras de dados com maior volume de documentos" (KURAMOTO, 2002).

Entretanto em 2002 já existiam algumas ferramentas que possibilitariam tal extração; essas ferramentas computacionais são conhecidas como analisadores. Sua função básica é identificar as classes gramaticais e os elementos sintáticos e semânticos que compõem uma sentença ou texto.

QUADRO 1 Analisadores língua portuguesa

FERRAMENTA	REFERÊNCIA	DIFICULDADES
VISL - Automatic Analysis of Portuguese	Bick, 1996[2]	(-) O envio do texto é feito pela web. (-) anotação manual: arquivo por arquivo. (-) Para anotações em um corpus maior é necessário adquirir licença. (-) Português europeu
Curupira - Parser para o português brasileiro	Martins, Hasegawa e Nunes, 2002[3]	(-) não está disponível ao público.
Grammar Play	Othero, 2004.	(-) léxico limitado. (-) apenas pequenas frases.
PoSiTagger	Aires, 2000.[4]	(-) dependente da ferramenta MXPost.
LX-Tagger / LX Suite	Natural Language and Speak group (NLX)[5]	(-) O envio do texto é feito pela web. (-) anotação manual: arquivo por arquivo. (-) Português europeu

Fonte: MARTINS; HASEGAWA; NUNES, 2002.

O trabalho de Miorelli (2001, p. 1) tem como objetivo "propor um método para a extração automática do Sintagma Nominal de sentenças em português, aplicando recursos de Processamento de Linguagem Natural (PLN) em Sistemas de Recuperação de Informações".

Entretanto, entre as diversas etapas do método ED-CER, detalhadas adiante, a etapa de etiquetagem, necessária para a extração dos SN, é realizada manualmente.

A autora sugere em suas considerações finais a utilização de um etiquetador que gere as etiquetas definidas em seu trabalho. E uma das funcionalidades da ferramenta Ogma é a aplicação de etiquetas para execução das regras adaptadas deste trabalho.

4 Escolha de descritores utilizando sintagmas nominais

Desde que a pesquisa de Kuramoto (1999; 2002) verificou a viabilidade do uso de sintagmas nominais em sistemas de recuperação de informação, muitos estudos se deram a partir desse, principalmente estudos sobre quais seriam os sintagmas que poderiam descrever e ser relevantes para o conteúdo do documento. A pesquisa efetuada por Souza (2005) trabalha sobre esse problema de escolha dos descritores, e propõe os seguintes passos:

extração dos sintagmas nominais do texto;

análise de cada um dos sintagmas nominais extraídos e cálculo da pontuação do mesmo como descritor.

[2] Mais informações: <http://visl.hum.sdu.dk/visl/pt>. 01 nov. 2008.
[3] Mais informações: <http://www.nilc.icmc.usp.br/nilc/tools/curupira.html>. 01 nov. 2008.
[4] Mais informações: <http://www.nilc.icmc.usp.br/nilc/projects/mestradorachel.html>. 01 nov. 2008.
[5] Mais informações: <http://nlx.di.fc.ul.pt/>. 01 nov. 2008.

Souza (2005) trabalhou com um corpus de 60 documentos: 30 artigos do periódico DataGramaZero e mais 30 artigos científicos da área de ciência da informação.

Toda a metodologia apresentada por Souza (2005) foi implementada na ferramenta OGMA e foi utilizada nos experimentos propostos por esta pesquisa. O OGMA realiza identificação da classe do sintagma nominal, bem como o cálculo da pontuação do mesmo como descritor de forma automática.

5 Similaridade de documentos eletrônicos

A classificação aplicada à prática da biblioteconomia corresponde a fornecer aos livros e documentos, de um modo geral, o lugar certo em um sistema de recuperação de informações, na qual existe uma coleção que abranja 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 a 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, objetivo desta pesquisa, tornam-se cada vez mais importantes num mundo com crescimento exponencial do volume de informação.

Na presente pesquisa utilizaram-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 aprimorar-se, obtendo resultados próximos de uma classificação feita pelo homem; essa busca por algumas técnicas inclui até o uso de inteligência artificial.

Conglomerados (ou *clustering)* correspondem às técnicas que permitam subdividir um conjunto de objetos em grupos. O objetivo é fazer com 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).

Para se localizar a similaridade entre dois documentos em um SRI utilizando *vector space model (VSM)*, calcula-se o cosseno do ângulo formado no vetor termo-por-documento. No *VSM* padrão quanto menor o ângulo, mais próximo de 1 será o cosseno, e mais similar será o documento em relação àquele termo.

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

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

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

1. um conjunto ordenado de documentos é recuperado, fornecendo uma melhor resposta à consulta;
2. documentos que têm mais termos em comum com a consulta tendem a ter maior similaridade;
3. aqueles termos com maiores pesos contribuem mais para a combinação do que os que têm menores pesos;
4. documentos maiores são favorecidos;
5. 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 esses documentos na matriz. Existem além do co-seno de similaridade outras medidas, sendo que a distancia euclidiana é também muito utilizada. A distância euclidiana entre dois documentos d1 e d2 é definida por:

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

Na qual: $w_{i,j}$ 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 válidas 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.

Naive Bayes é o método de classificação baseado em inferência bayesiana. Trabalha com dados contínuos e discretos. Para dados discretos os valores de probabilidades são coletados através da contagem nos grupos dos documentos. Para dados contínuos, ele assume que os valores sigam uma função de distribuição normal, assim, as probabilidades são inferidas a partir da média e do desvio padrão de grupos dos documentos.

O algoritmo K-means (ou K-médias) tem o objetivo de fornecer um agrupamento de objetos de acordo com os seus próprios dados. Essa classificação é baseada em análise e comparações entre os valores numéricos dos dados fornecidos. Dessa maneira, o algoritmo realiza um agrupamento automático sem a necessidade de nenhuma supervisão humana, ou seja, sem nenhum pré-agrupamento existente. Por causa

desta característica, o *K-means* é considerado um algoritmo de mineração de dados não supervisionado.

O algoritmo analisa todas as instâncias fornecidas e as agrupa, isto é, o algoritmo vai indicar uma classe (*cluster*) e vai dizer que linhas pertencem a essa classe. O usuário fornece ao algoritmo a quantidade de classes que ele deseja (k). Este número de classes que deve ser passada para o algoritmo é chamado de *k* de onde vem a primeira letra do algoritmo: *K-means*.

6 Instrumentos

Nesta pesquisa, optou-se pelo desenvolvimento de uma ferramenta própria devido às limitações das ferramentas disponíveis para a extração de sintagmas nominais. O OGMA possibilitou a aplicação do experimento proposto na metodologia, destacando a possibilidade de análise de um *corpus* maior devido à total automatização.

O nome Ogma foi dado em homenagem ao deus celta Ogma (nome reduzido de *Ogmious*). Esse deus criou mecanismos de linguagem e engrandeceu a comunicação do povo celta.

O aplicativo foi desenvolvido na ferramenta *visual studio .NET* em linguagem C#. Por se tratar de uma ferramenta para análise de texto optou-se pelo desenvolvimento em modo texto, o que não impede que sejam desenvolvidas interfaces gráficas posteriormente.

O primeiro desafio na construção dessa ferramenta foi a elaboração de um léxico da língua portuguesa completo o suficiente para permitir análises e conseqüente etiquetagem do texto.

A primeira etapa para a construção desse dicionário foi a adaptação de arquivos com o vocabulário utilizado pelo BR/ISPELL[6]. Essa ferramenta foi desenvolvida para verificar a ortografia de projetos de código aberto. Através da adaptação desses arquivos foi possível a construção de um arquivo de dados (optou-se pelo uso do *access*) com uma tabela de 41978 nomes e adjetivos.

Outro item necessário era a identificação dos verbos. Utilizando a ferramenta conjugue[7] e uma base de dados de 5000 verbos conseguiram-se reunir em outra tabela do banco de dados 292.720 verbos. Devido a regras de identificação de sintagmas nominais utilizadas nesta pesquisa também foi necessário identificar os verbos no particípio; esses verbos receberam uma identificação diferenciada 'VP' no lugar de 'VB'.

Finalmente, através de um processo manual de digitação, tendo como base a gramática de Tufano (1990), conseguiram-se reunir 475 palavras de diversas classes gramaticais. Essas palavras são as mesmas que foram utilizadas para compor a lista de *stopwords*.

[6] Para mais informações: <http://www.ime.usp.br/~ueda/br.ispell>. 01 nov. 2008.
[7] O conjugue é um programa para linux capaz de conjugar verbos da língua portuguesa, a partir de um banco de regras pré-definidas de conjugação.

O programa Weka - *Waikato environment for knowledge analysis* começou a ser escrito em 1993, usando Java, na Universidade de Wakato localizada na Nova Zelândia.

A licença utilizada pelo Weka é a *General public license* – GPL, sendo o pacote Weka formado por um conjunto de implementações de algoritmos de diversas técnicas de mineração de dados e textos.

O Weka agrega diversos algoritmos provenientes de diferentes abordagens/paradigmas na subárea da inteligência artificial dedicada ao estudo da aprendizagem por parte de máquinas. Entre esses, algoritmos de classificação e agrupamento.

Nesta pesquisa, foram utilizados dois *corpora*:

O primeiro, constituído de 50 artigos selecionados do Encontro Nacional de Pesquisa em Ciência da Informação - ENANCIB 2005. Nesse congresso, os trabalhos apresentados foram divididos em grupos de trabalho (GT) de acordo com o assunto tratado no artigo.

Dessa forma, selecionaram-se aleatoriamente 10 documentos de 5 grupos de trabalhos – GT.

Foram selecionados os cinco primeiros grupos:

GT 1: Estudos históricos e epistemológicos da informação
GT 2: Organização do conhecimento e representação da informação
GT 3: Mediação, circulação e uso da informação
GT 4: Gestão de unidades de informação
GT 5: Política, ética e economia da informação

Os artigos foram convertidos do formato original PDF em arquivos texto pela ferramenta *Adobe Acrobat*. Os arquivos resultantes (formato texto) ficaram entre 21 kbytes, 3156 termos (texto 9 do GT 5) e 213 kbytes, 29857 termos (texto 10 do GT 2).

A esse *corpus* deu-se o nome de "ENANCIB05".

Do total de 100492 termos, 22320 são distintos e 6642 (7%) são únicos.

Também se optou por utilizar um segundo *corpus*, formado por textos menores e de conteúdo jornalístico. O objetivo foi avaliar o comportamento em um *corpus* diferente e com conteúdo bem definido em relação aos assuntos.

Para isso extraiu-se do site do jornal Hoje em Dia todas as notícias de 2004. Em seguida realizou-se um trabalho de classificação manual, trocando o nome dos arquivos de acordo com o caderno do qual o texto foi retirado. Após esse processo, realizou-se seleção aleatória de 40 notícias dos seguintes cadernos: Informática, Turismo e Veículos.

Para compor esse *corpus* e dificultar a etapa do agrupamento automático, decidiu-se adicionar mais um tema. Utilizou-se outro corpus que possuísse essa separação, por temas, disponibilizado pela linguateca TeMario[8].

[8] Para mais informações: <http://www.linguateca.pt/Repositorio/TeMario/>. Acesso em: 01 nov. 2008.

Do TeMario foram retirados os textos do Jornal do Brasil, seção Internacional (Internacional 20 textos, 12.098 termos e média de termos 604) e Folha de São Paulo, seção Mundo (20 textos, 13.739 termos e média de termos 686) para compor esta nova temática.

O novo *corpus* denominado "JORNAL04", segundo deste experimento, ficou então com 160 textos divididos entre quatro temas: Informática, Turismo, Veículos e Mundo.

7 Experimento

O experimento foi aplicado nos dois *corpora* e constituiu-se das seguintes etapas:

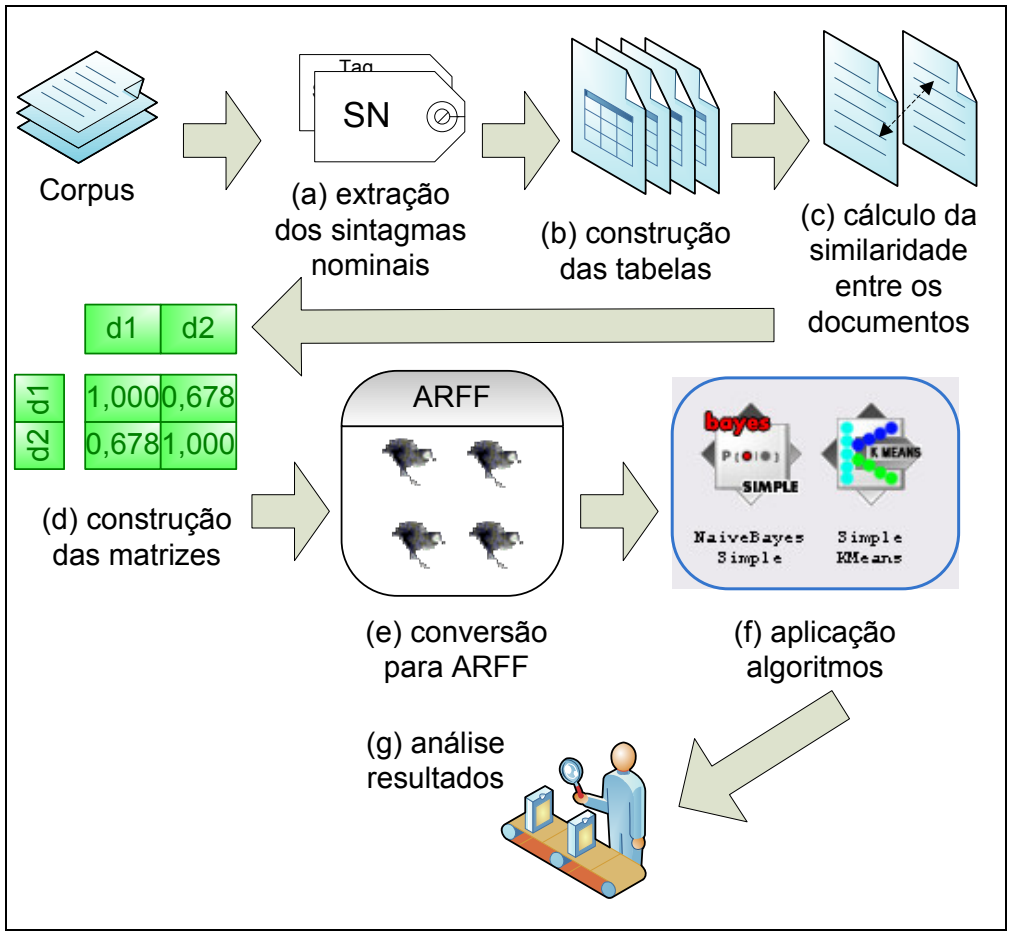

FIGURA 1 – Etapas do experimento consolidado

Fonte: Dados da pesquisa.

a) extração dos sintagmas nominais dos textos analisados;
b) construção das tabelas de descritores com os pesos por: termo, termos sem *stopwords*, sintagmas nominais e sintagmas nominais pontuados. No *corpus* JORNAIS04 utilizaram-se também os sintagmas nominais aninhados e sintagmas nominais aninhados pontuados;

c) cálculo do índice de similaridade entre as tabelas de cada arquivo;
d) criação de um arquivo para cada método, contendo a matriz com os resultados encontrados no item anterior. A matriz terá o número de linhas e o número de colunas igual ao número de documentos;
e) conversão dessas tabelas para o formato Arff do *software* Weka;
f) aplicação dos algoritmos *Naive Bayes* e *Simplekmeans;*
g) análise dos resultados.

Para o experimento consolidado, utilizou-se, além do *corpus* ENANCIB05, do *corpus* JORNAIS04. Conforme apontado nos resultados obtidos no experimento prospectivo, verificou-se a necessidade de um *corpus* com mais documentos e com temas mais definidos.

Também após a aplicação do experimento prospectivo, verificou-se a necessidade de se aprimorar a metodologia da análise dos resultados e, para isso, utilizou-se o *software* WEKA, gerando-se uma análise com um número maior de indicadores quantitativos.

8 Resultados

Durante toda a apresentação dos dados, utilizaram-se as seguintes nomenclaturas para definir o método de análise (QUADRO 2):

QUADRO 2 - Siglas dos métodos de análise utilizados nos experimentos

SIGLA	MÉTODO	DESCRIÇÃO
TT	Termo	As tabelas de descritores contêm todas as palavras do documento e seus respectivos pesos.
TTS	Termo sem *stopwords*	A tabela de descritores é construída com base em todas as palavras do documento, com exceção das presnetes na lista de *stopwords*.
TC	Sintagmas Nominais (máximos)	A tabela de descritores é construída com base nos sintagmas nominais extraídos de cada documento.
TR	Sintagmas Nominais (máximos) Pontuados	A tabela de descritores é construída de acordo com o cálculo realizado da pontuação como descritor de cada sintagma nominal.
TCA	Sintagmas Nominais (incluindo os aninhados)	A tabela de descritores é construída com base nos sintagmas nominais máximos e aninhados extraídos de cada documento.
TRA	Sintagmas Nominais (incluindo os aninhados) Pontuados	A tabela de descritores é construída de acordo com o cálculo realizado da pontuação como descritor de cada sintagma nominal máximo e aninhado.

Fonte: Dados da pesquisa.

Os documentos dos *corpora* ENANCIB05 e JORNAIS04 foram comparados entre si, utilizando os métodos propostos.

A aplicação dos algoritmos de *Naive Bayes* e *simplekmeans* através do WEKA, nesses dados, apresentaram os seguintes resultados consolidados na TAB. 1 para o *corpus* do ENANCIB05:

TABELA 1- Resultado Weka para o *corpus* ENANCIB05

Método	***Naive Bayes***		***SimpleKMeans***	
	Classificados corretamente	%	Agrupados corretamente	%
TTS - Termos sem *stopwords*	26/50	52%	22/50	44%
TT – Termos	25/50	50%	21/50	42%
TR - Sintagmas Nominais Pontuados	23/50	46%	20/50	40%
TC - Sintagmas Nominais	12/50	24%	15/50	30%

Fonte: Dados da pesquisa.

Os dados da TAB. 1 estão representados na FIG. 2, abaixo:

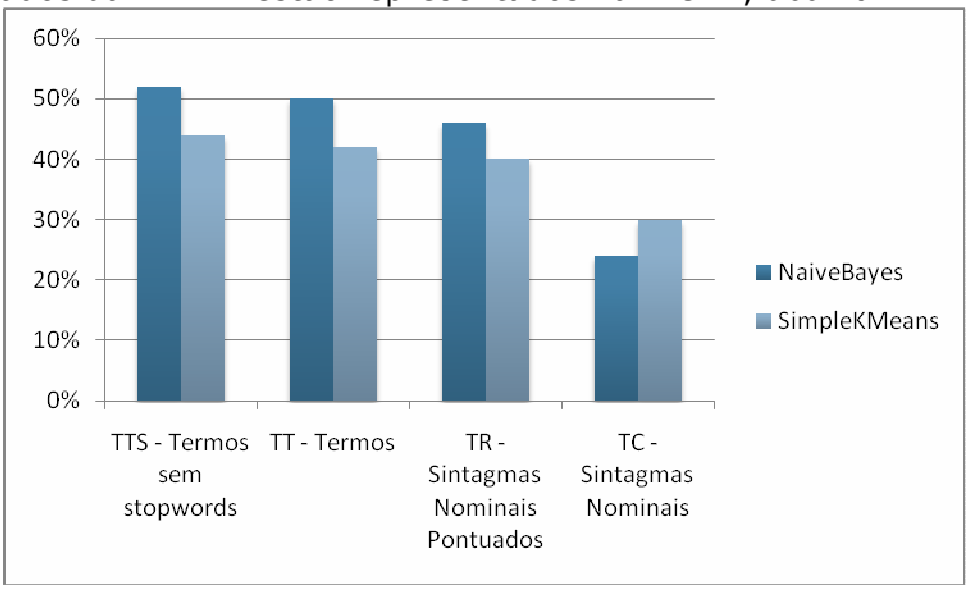

FIGURA 2 – Resultados Weka para o corpus ENANCIB05

Fonte: Dados da pesquisa.

Mesmo utilizando-se os algoritmos da ferramenta WEKA no corpus ENANCIB05, não foi possível obter conclusões sobre qual o melhor método, visto que a diferença do número de documentos classificados corretamente entre os métodos de termos e sintagmas nominais pontuados foi muito pequena. Apenas o método de sintagmas nominais simples apresentou resultados inferiores aos outros métodos.

Os resultados foram bem semelhantes aos encontrados no experimento prospectivo. Entretanto, o uso da técnica de *cross validation* para seleção do conjunto de treinamento utilizada pelo WEKA fez com que o método que usou termos sem *stopwords* tivesse um resultado um pouco

melhor (52%) que no uso de termos apenas (%50), conforme a TAB. 2. No experimento prospectivo, o corpus de treinamento era fixo, composto dos 5 primeiros artigos de cada GT.

TABELA 2 - Comparação experimento prospectivo x experimento consolidado

	Experimento prospectivo		**Experimento consolidado (*Naive Bayes*)**	
Método	**Doc. corretos**	**%**	**Doc. corretos**	**%**
TT – Termos	14/25	56%	25/50	50%
TS - Termos sem *stopwords*	12/25	48%	26/50	52%
TR - Sintagmas Nominais Pontuados	11/25	44%	23/50	46%
TC - Sintagmas Nominais	6/25	24%	12/50	24%

Fonte: Dados da pesquisa.

Para o corpus JORNAIS04, foram utilizados, além dos quatro métodos anteriormente descritos (termos - TT, termos sem stopwords - TS, sintagmas nominais – TC e sintagmas nominais pontuados - TR), mais dois: um considerando os sintagmas nominais aninhados (TCA) e outro considerando os sintagmas nominais aninhados pontuados (TRA).

Para cada documento da coleção, gerou-se uma tabela com a lista de descritores utilizados para a indexação do documento e seus respectivos pesos. Essa tabela foi referência na comparação entre os documentos.

Na TAB. 3 é apresentado o número médio de descritores das tabelas geradas a partir dos documentos do *corpus* JORNAIS04.

TABELA 3 -Número médio de descritores por documento e método, corpus JORNAIS04

Método	**Número médio de descritores por documento**
TT – Termos	331
TTS – Termos sem *stopwords*	296
TC – Sintagmas Nominais	160
TR – Sintagmas Nominais pontuados	160
TCA – SN Aninhados	245
TRA – SN Aninhados e pontuados	245

Fonte: Dados da pesquisa.

O experimento, utilizando o WEKA algoritmo classificador *Naive Bayes,* foi repetido com o corpus JORNAIS04 e o resultado consolidado se encontra na TAB. 4.

TABELA 4 -Resultado Weka/*Naive Bayes*, corpus JORNAIS04

Corpus: JORNAIS04		
Método	***Naive Bayes***	
	Classificados corretamente	%
TS - Termos sem *stopwords*	147/160	91%
TC - Sintagmas Nominais	147/160	91%
TRA - SN Aninhados Pontuados	137/160	85%
TCA - SN Aninhados	136/160	85%
TR - Sintagmas Nominais Pontuados	132/160	82%
TT - Termos	106/160	66%

Fonte: Dados da pesquisa.

O arquivo com o quadro comparativo de cada método também foi submetido ao algoritmo *SimpleKMeans* do WEKA. Os resultados encontram-se na TAB. 5:

TABELA 5- Resultado Weka/*SimpleKMeans*, corpus JORNAIS04

Corpus: JORNAIS04		
Método	***SimpleKMeans***	
	Agrupados corretamente	%
TRA - SN Aninhados Pontuados	129/160	81%
TS - Termos sem *stopwords*	126/160	79%
TCA - SN Aninhados	109/160	68%
TC - Sintagmas Nominais	89/160	56%
TT - Termos	80/160	50%
TR - Sintagmas Nominais Pontuados	66/160	43%

Fonte Fonte: Dados da pesquisa.

Observa-se que os melhores resultados foram obtidos utilizando-se a comparação entre tabelas de descritores contendo a relação termos sem *stopwords* (TTS) e as tabelas de sintagmas nominais aninhados e pontuados (TRA), utilizando-se o algoritmo de classificação *Naive Bayes*. Ambos classificaram 147 documentos corretamente, num total de 91% do *corpus*.

Os resultados atingidos pelo *simplekmeans,* para agrupamento, foram inferiores ao classificador *Naive Bayes,* em ambos os experimentos.

Os resultados dos experimentos apontam que os métodos que envolveram o uso de sintagmas nominais na classificação automatizada de documentos apresentaram índices semelhantes ao dos termos sem *stopwords*. Por exemplo, no corpus JORNAIS04, os sintagmas nominais e os termos sem *stopwords* atingiram o mesmo resultado com o *Naive Bayes*, obtendo 91% (147 documentos) de classificação correta. No algoritmo de cluster *simplekmeans,* o sintagma nominal aninhado e pontuado se apresentou *um* pouco melhor que o termo sem *stopwords;* classificando corretamente 129 documentos (81%) contra 126 (79%) dos termos sem *stopwords*.

Apesar de resultados similares, o uso de sintagmas nominais envolveu um processamento computacional muito maior que o uso de termos sem *stopwords*. Como visto, para o computador extrair os sintagmas nominais foi necessário todo um processo de etiquetagem (anotação) e aplicação de regras gramaticais ao documento, enquanto que no uso do método de termos sem *stopwords* o processo foi bem mais simples. Dessa forma, a relação custo *versus* benefício de se utilizarem sintagmas nominais na classificação e no agrupamento de documentos eletrônicos não se apresentou interessante.

Em contraponto, os piores resultados de classificação e agrupamento ficaram para o uso de todas as palavras (termos) que compõe o documento. Isto pode indicar a importância de se submeter o documento a um tratamento prévio antes da elaboração da tabela de termos e pesos, mesmo que seja somente a retirada das *stopwords*.

9 Considerações finais

Os documentos utilizados nos *corpora* deste artigo foram elaborados através da linguagem natural. Essa linguagem, em alguns casos, foi ambígua semanticamente e apresentou problemas decorrentes da diferença do vocabulário usado nos textos. Os métodos de descoberta de conglomerados, ou classificação, mostraram-se extremamente dependentes de técnicas de pré-processamento dos textos que visassem à padronizá-los, minimizando os problemas do vocabulário e representando seu conteúdo de forma mais correta e fácil de ser trabalhada pela máquina.

A presente pesquisa focou justamente esse ponto, buscando demonstrar que a utilização de sintagmas nominais é capaz de representar o conteúdo dos documentos, servindo como descritores ou características para o processo de classificação ou descoberta de conglomerados, melhorando a precisão desse processo.

Este trabalho representa uma continuidade de estudos nas áreas de ciência da informação, computação e lingüística e novos trabalhos podem, a partir dos resultados alcançados aqui, elaborar novos experimentos. A ferramenta OGMA, a primeira ferramenta específica para extração de sintagmas nominais em português, estará disponível para download no endereço http://www.luizmaia.com.br/ogma.

Referências

AIRES, R. V. X. *Implementação, adaptação, combinação e avaliação de etiquetadores para o português do Brasil*. 2000. Dissertação (Mestrado em Ciência da Computação e Matemática Computacional) – Instituto de Ciências Matemáticas, Universidade de São Paulo, São Paulo, 2000.

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

BICK. Automatic parsing of portuguese. In: ENCONTRO PARA O PROCESSAMENTO COMPUTACIONAL DE PORTUGUÊS ESCRITO E FALADO, 2., Curitiva, 1996. *Anais…* Curitiva: CEFER- PR, 1996.

CHOMSKY, N. *Diálogos com Mitsou Ronat.* São Paulo: Cultrix, 1969.

CROFT, W. B. *Advances in information retrieval.* London: Academic Publishers, 2000.

FRANTZ, V.; SHAPIRO, J.; VOISKUNSKII, V. *Automated information retrieval:* theory and methods. San Diego, CA: Academic Press, 1997. 365 p.

JANSSENS, F. *Clustering of scientific fields by integrating text Mining and bibliometrics*. [s.l.]: Katholieke Universiteit Leuven; Faculteit Ingenieurswetenschappen, 2007.

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

_____. Sintagmas Nominais: uma nova proposta para a Recuperação da Informação. *DataGramaZero*, v. 3, n. 1, fev. 2002. Disponível em: <http://www.dgz.org.br/fev02/Art_03.htm>. Acesso em: 01 nov. 2008.

LOPES, E. *Fundamentos de lingüística contemporânea.* São Paulo: Cultrix, 1999.

MAIA, Luiz Cláudio Gomes. *Uso de sintagmas nominais na classificação automática de documentos eletrônicos.* 2008. Tese (Doutorado em Ciência da Informação) – Universidade Federal de Minas Gerais – UFMG. Belo Horizonte, 2008.

MARTINS, R. T.; HASEGAWA, R.; NUNES, M. G. V. *Curupira*: um parser functional para o português. [s.l.: s. n.], 2002. Relatório.

MEADOW, C.T.; BOYCE, B. R.; KRAFT, D. H. *Text Information Retrieval Systems.* San Diego: Academic Press, 2000. 364 p.

MIORELLI, S. T. *ED-CER*: extração do sintagma nominal em sentenças em português. 2001. Dissertação (Mestrado em Ciência da Computação) – Faculdade de Informática, Pontifícia Universidade Católica do Rio Grande do Sul, Porto Alegre, 2001.

OTHERO, G. A. *Grammar play*: um parser sintático em Prolog para a lingual portuguesa. 2004. Dissertação (Mestrado em Letras) – Pontificia Universidade Católica do Rio Grande do Sul, Porto Alegre, 2004.

PERINI M, A. O Sintagma nominal em português: estrutura, significado e função, Revista de Estudos da Linguagem, n. esp., 1996. Disponível em: <http://relin.letras.ufmg.br/revista/upload/Relin_NEspecial_1996.pdf>. Acesso em: 01 nov. 2008.

POLANCO, X.; FRANÇOIS, C. Data clustering and cluster mapping or visualization in text processing and mining. In: INTERNATIONAL ISKO CONFERENCE, 6., 2000, Toronto. *Proceedings…*Toronto: Ergon Verlag: Würzburg, 2000. p. 359-365.

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

SILVA, M. C. P. S.; KOCH, I. V. *Lingüística aplicada ao português*: sintaxe. 14 ed. São Paulo: Cortez, 2007.

SOUZA, R. R. *Uma proposta de metodologia para escolha automática de descritores utilizando sintagmas nominais.* 2005. Tese (Doutorado em Ciência da Informação) – Escola de Ciência da Informação, Universidade Federal de Minas Gerais, Belo Horizonte, 2005.

TRASK, R. L. *Dicionário de linguagem e lingüística.* São Paulo: Contexto, 2004.

TUFANO, D. *Estudos de língua portuguesa*: gramática. 2 ed. São Paulo: Moderna, 1990.