Busca Guiada de Patentes de Bioinformática

Marcio Branquinho Dutra[1], José Antonio Camacho-Guerrero[2], José Augusto Baranauskas[1], Alessandra Alaniz Macedo[1]

[1]Departamento de Computação e Matemática (DCM), Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto (FFCLRP-USP), Av. dos Bandeirantes, 3900 – Campus da USP – Ribeirão Preto-SP – Brasil
mbdutra@usp.br, augusto@usp.br, ale.alaniz@usp.br
[2]i-MedSys - Innovative Medical Informatics, http://www.i-medsys.com/, Rua Magda Perona Frossard, 750 – Ribeirão Preto-SP – Brasil
jose.camacho@imedsys.com.br

1

Abstract. *Recent researches have suggested that patents gather about 70% of world knowledge and offer more detailed information than scientific papers. However, patents are not wide used by academic community as reference sources, because of limitations on public search tools. The patent's language is complex and that is harder on multidisciplinary fields, such as bioinformatics. Recent studies have indicated the use of classifications, linguistic artifacts and relevance feedback are important mechanisms to improve search results. This paper presents a classifier as a filter of a guided patent search system.*

Resumo. *Pesquisas recentes demonstram que as patentes armazenam em torno de 70% do conhecimento mundial. No entanto, esses documentos são pouco utilizados como fontes de referência no meio acadêmico por serem considerados pouco acessíveis. As ferramentas públicas de busca são limitadas, a linguagem utilizada nas redações é complexa e para patentes de áreas multidisciplinares, como a bioinformática, a complexidade é maior. Estudos recentes demonstram que a utilização de mecanismos como classificadores e artefatos linguísticos auxiliam na obtenção de resultados mais relevantes aos usuários. Este artigo apresenta um classificador de documentos de patentes, que atua como filtro no sistema de busca guiada de patentes de bioinformática.*

1. Introdução

Patentes são licenças públicas temporárias outorgadas pelo Estado aos titulares para a exploração econômica de suas invenções [Barbosa 2003]. Essas invenções são publicadas pelos escritórios e descrevem todos os detalhes técnicos, a fim de divulgar os avanços tecnológicos atingidos [Barbosa 2003]. De acordo com [WIPO 2010b], entre os países com maior número de registros estão Estados Unidos, Japão, Alemanha e Coreia do Sul. Nesse contexto, o Brasil, apesar do crescimento econômico dos últimos anos, ainda apresenta números consideravelmente inferiores aos outros países emergentes como China, Rússia e Índia [FAPESP 2010].

Atualmente, todos os escritórios internacionais têm sido criticados mundialmente devido ao tempo para a análise de cada pedido de patente [INPI 2011a] [Rodriguez 2010] [Reid 2011]. No Brasil o INPI despende oito anos para emitir o parecer final [INPI 2011b], enquanto em órgãos, como o americano USPTO (*United States Patent and Trademark Office*), cada pedido é finalizado, em média, em quatro anos [USPTO 2012]. Essa demora no Brasil e no exterior é devido à quantidade de pedidos depositados, à quantidade de examinadores, à quantidade e qualidade de ferramentas computacionais que auxiliam nas análises e buscas por patentes similares e à complexidade do processo [USPTO 2012] [EPO 2011][INPI 2011b].

Pesquisas recentes demonstram que as patentes armazenam em torno de 70% do conhecimento mundial, disponibilizam informações técnicas mais detalhadas que artigos científicos, auxiliam na divulgação do conhecimento e estimulam o desenvolvimento tecnológico de diversas áreas [Lupu et al. 2011]. No entanto, as patentes ainda são pouco utilizadas como fontes de referência no meio acadêmico por serem consideradas pouco acessíveis. As ferramentas públicas de busca trabalham principalmente com pesquisas por palavra-chave e, segundo [Eisinger et al. 2012] e [Park 2012], a busca por patentes deve ir além dessa técnica, pois a linguagem utilizada nas redações de patentes é complexa e com poucas descrições específicas para tornar a patente o mais abrangente possível.

O presente trabalho investiga a criação de uma máquina de busca por patentes em bioinformática guiada por técnicas de processamento de linguagem natural, recuperação de informação e *machine learning* na tentativa de facilitar a busca por patentes relacionadas, diminuindo a sobrecarga cognitiva e agilizando o processo. Este artigo apresenta o classificador de documentos textuais de registro de patentes, que atua como filtro no sistema de busca guiada de patentes. Futuramente, espera-se estender o domínio de aplicação do sistema para outras áreas de saúde e de ciências biomédicas, além de bioinformática.

O restante deste artigo está organizado da seguinte forma: a Seção 2 descreve trabalhos relacionados, a Seção 3 apresenta classificações de patentes, a Seção 4 descreve o módulo de classificação de patente dentro do sistema de busca guiada de patentes, a Seção 5 ilustra as experimentações de classificações, e finalmente a Seção 6 apresenta conclusões e trabalhos futuros.

2. Trabalhos Relacionados

Em [Eisinger et al. 2012], os autores fazem uma análise comparativa da utilização de termos MeSH e informações de classificação da taxonomia IPC para aumentar a recuperação de documentos relevantes em buscas por patentes biomédicas. O trabalho encontra similaridades nas estruturas dos dois artefatos, mas aponta o IPC como mais complexo e menos acessível por utilizar muitos códigos. Enquanto isso, o MeSH utiliza termos inteligíveis, o que facilita sua utilização por usuários menos especializados. Neste artigo, os autores propõem alternativas para expandir consultas baseadas em palavras-chave.

[Park 2012] efetuam análise comparativa de buscas por patentes de bioinformática utilizando apenas palavras-chave e palavras-chave associadas à informações de classificação. Os autores utilizaram patentes disponibilizadas pelo escritório coreano

702	Processamento de dados: medições, calibrações ou testes			**Bioinformatics**
	1	Sistemas de medição em ambiente específico		**Classes**
		19	Biológico ou Bioquímico	***702/19***
			20 Determinação de sequência de genes	***702/20***
			21 *Contagem de células ou análise da forma ou tamanho*	***702/21***
703	Processamento de dados: projeto estrutural, modelagem, simulação e emulação			
	6	Simulação de sistemas ou dispositivos não-elétricos		
		11	Biológico ou Bioquímico	**703/11**
		12	Químico	**703/12**

Tabela 1. Classes USPC de bioinformática e hierarquias [USPTO 2012].

(KIPO[1]) através do sistema KIPRIS[2]. Para identificar as classes IPC de bioinformática foi utilizado o sistema USPC-to-IPC *reverse concordance system*[3] de correspondência entre USPC e IPC. Os autores apontam que a utilização de informações de classificação aumentou a quantidade de documentos relevantes recuperados e propõem análises mais detalhadas em trabalhos futuros.

[Teixeira et al. 2012] propõem um método de indexação automática de artigos científicos da área de informática em saúde (IS) utilizando o algoritmo de classificação *Multinomial Naive Bayes*. O autores apontam que a grande quantidade de publicações da área e a característica interdisciplinar da IS exigem ferramentas computacionais eficientes que auxiliem na indexação e recuperação de documentos. Os resultados encontrados foram considerados satisfatórios na indexação de artigos de IS. Os autores sugerem como uma das atividades futuras, a aplicação de técnicas de redução de dimensionalidade visando melhorar o desempenho do classificador utilizado.

Em [Mukherjea and Bamba 2004], no contexto biomédico as ferramentas gratuitas atuais, baseadas em técnicas de pesquisa por palavra-chave e modelo booleano de recuperação de informação não se mostram eficientes no relacionamento e na recuperação de informações relevantes aos usuários. Esse fato provoca a busca por registros na intersecção de várias classes nos clássicos sistemas de classificações de patentes. Na maioria dos casos de busca, é essencial a manipulação manual das buscas e dos resultados em quantidades crescentes de documentos. Esse cenário caracteriza dois problemas típicos de busca: sobrecarga cognitiva e gasto excessivo de tempo para buscar e filtrar informação.

3. Classificação de Patentes

De acordo com [Lupu et al. 2011], uma maneira eficiente de facilitar a busca de objetos é organizá-los em grupos com características similares. Por exemplo, uma biblioteca organiza todos os livros por assunto para facilitar a busca por livros específicos e permitir encontrar mais informação sobre o mesmo tópico ou similares.

Patentes depositadas nos escritórios internacionais são classificadas de acordo com algumas taxonomias, sendo as mais difundidas a internacional *Internation Patent Classification* (IPC), a americana *United States Patent Classification* (USPC), a europeia *European Patent Classification* (ECLA) e a japonesa *Japanese File Index and F-Term*

[1]http://www.kipo.go.kr/en
[2]http://www.kipris.or.kr/enghome
[3]http://www.uspto.gov/web/patents/classification/international/ipc/ipc8/ipc_concordance/ipcsel.htm

(FI/F-Term). A taxonomia internacional (IPC) e a americana (USPC) apresentam classes específicas para a área de bioinformática. A classificação específica americana foi criada em 1999 [USPTO 2012], enquanto a IPC é mais recente e foi criada em 2010 [WIPO 2010a]. O presente trabalho aborda a classificação americana, uma vez que a coleção utilizada nos experimentos foi obtida no site do USPTO. A Tabela 1 detalha as subclasses e a hierarquia.

Existem dois tipos de classificação USPC: obrigatória (*mandatory*) e arbitrária (*discretionary*). De acordo com [USPTO 2005], cada reivindicação de uma patente deve receber uma classificação para que o documento seja rastreável por pesquisas classificadas. Dessa maneira, todas as patentes recebem, ao menos, uma classificação obrigatória. Classificações obrigatórias são subdivididas em classificação original, do inglês *original classification* (OR) e classificações de referência cruzada, do inglês *cross-reference* (XR). A OR identifica o documento e é obtida da classificação atribuída à reivindicação principal da patente. As XR são obtidas das classificações atribuídas às reivindicações secundárias. As classificações do tipo arbitrária (XD) também são de referência cruzada e são atribuídas às patentes pelas unidades de análise, que desejam adicionar informações específicas que auxiliam nas pesquisas classificadas.

4. Proposta de Busca Guiada de Patentes

O presente trabalho investiga a criação de uma máquina de busca por patentes em bioinformática. As consultas de usuários efetuadas nesse sistema deverão ser guiadas por técnicas de processamento de linguagem natural, recuperação de informação e *machine learning* de modo a retornarem resultados mais completos e relevantes aos usuários para que o mesmo não necessite filtrar resultados e realizar novas buscas. Conforme demonstrado na Figura 1, propõe-se neste artigo um filtro, modelado segundo um processo de descoberta de conhecimento (KDD - Knowledge-Discovery in Databases), para guiar as buscas no sistema de patentes. O filtro processa as consultas de usuários e apresenta, por meio de um *ranking*, as classes de patentes de bioinformática relacionadas à consulta do usuário. Em seguida, o módulo de *relacionamento* calcula a similaridade da consulta original com cada classe de patentes do *ranking* elaborado pelo filtro. Esse procedimento de visualização, via relacionamentos, visa orientar e diminuir o campo de busca, de maneira que sejam encontrados mais documentos relevantes.

O processo de descoberta de conhecimento para o sistema de busca demanda um processo com etapas bem definidas que transformam os dados brutos de registro de patentes em informações relevantes aos usuários. Conforme ilustra a Figura 2, o modelo utilizado na proposta é composto por cinco etapas: (1) Seleção, (2) Pré-processamento, (3) Transformação, (4) Mineração de Texto e (5) Avaliação.

Na *Etapa de Seleção (1)*, foram identificadas as classes USPC que categorizam as tecnologias de bioinformática e a distribuição das classes na coleção utilizada. Após a análise do corpus, o processo de coleta foi iniciado para capturar a visão lógica de cada documento. Primeiramente identificou-se a estrutura dos documentos da coleção, em seguida definiu-se os campos a serem coletados.

Durante a coleta, a *Etapa de Pré-processamento (2)* dos dados foi iniciada, a qual corresponde às tarefas de eliminação de caracteres especiais e *stopwords* e a redução de cada termo ao seu radical (*stemming*). As tarefas de coleta e de pré-processamento, bem

como os pesos de cada termo, são descritos na Subseção 5.1.

Na *Etapa Transformação (3)* de atributos no processo de descoberta de conhecimento é efetuada para padronizar os valores, de maneira a normalizá-los e adaptá-los ao classificador que será aplicado. Nessa etapa são calculados os pesos de cada termo a partir de suas frequências. O método adotado para o cálculo de pesos foi o *tf*idf*.

De acordo com [Baeza-Yates and Ribeiro-Neto 1999], esse modelo se mostra interessante pois, para o cálculo dos pesos de cada termo, considera dois componentes importantes em uma coleção. O primeiro componente *tf* ou *term frequency* é dado pela simples frequência de cada termo em um documento e denota a significância desse termo para o documento. O segundo componente *idf* (*inverse document frequency*) é dado pelo inverso da frequência de um termo em todos os documentos da coleção. Segundo os mesmos autores, pesos compostos por esses dois fatores demonstram a representatividade de cada termo perante cada documento e perante a coleção como um todo.

Na *Etapa de Mineração de Textos (4)*, foram utilizados três algoritmos de classificação: Naive Bayes Multinomial (NBM), k-Nearest Neighbors (kNN) e *Sequential Minimal Optimization* (SMO), o qual resolve o problema de Máquinas de Vetor Suporte, do inglês, *Support Vector Machines*(SVM).

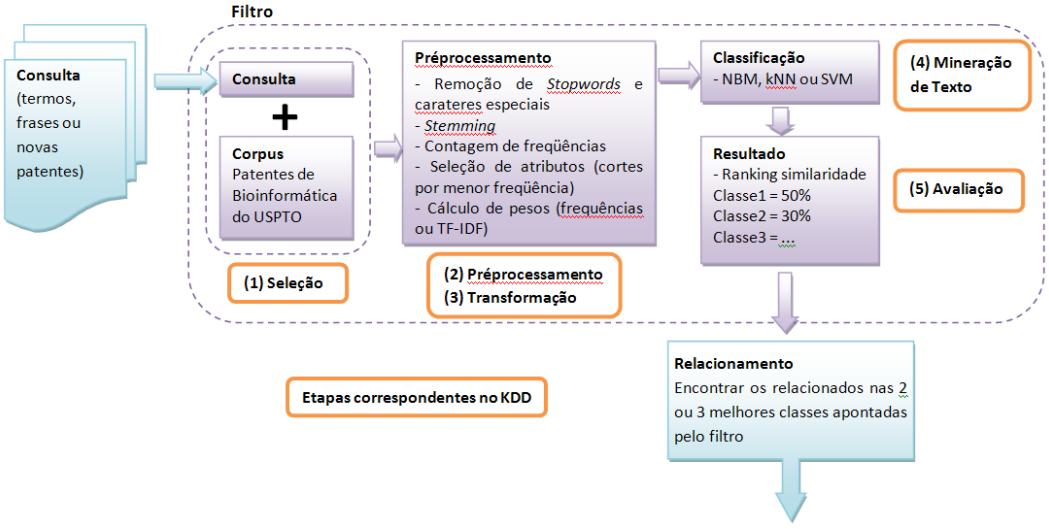

Figura 1. Visão geral do sistema de busca guiada de patentes de bioinformática.

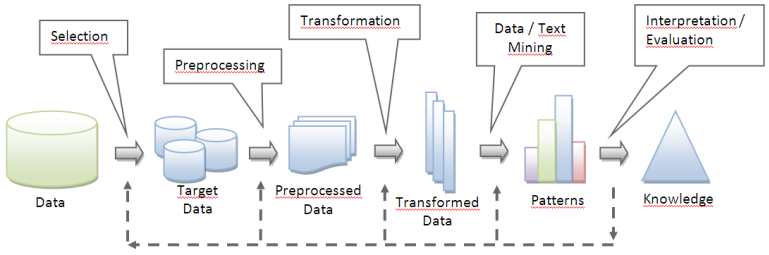

Figura 2. Processo de descoberta de conhecimento do sistema de busca guiada. [Tan et al. 2006]

O Naive Bayes Multinomial é um algoritmo probabilístico de complexidade computacional linear, baseado no Teorema de Bayes e consiste em uma variação do Naive Bayes (NB). Enquanto o original NB trabalha apenas com atributos categóricos, o Multinomial NB permite processar instâncias com atributos numéricos. Devido à essas características, esse algoritmo é comumente utilizado na classificação de textos, especialmente sob o modelo *bag of words*, o qual representa cada documento por termos de índice independentes. Esse classificador calcula a probabilidade de cada conjunto de termos de índice relativa a cada classe e define uma distribuição multinomial para cada documento por um vetor de probabilidades. O algoritmo NB também é utilizado na classificação de patentes [Lupu et al. 2011].

O k-NN é um dos mais simples e bem difundidos algoritmos de classificação baseado em instâncias. Esse algoritmo compara cada nova instância a ser classificada com os documentos de sua coleção de treinamento. A classe que será atribuída ao novo documento é definida pelos k documentos mais próximos (similares). De acordo com [Lupu et al. 2011], esse algoritmo apresenta bom desempenho com um número reduzido de classes, como é o caso do presente trabalho para as classes de bioinformática. Além disso, o kNN é utilizado comumente em pesquisas pelo estado da arte de tecnologias.

O SVM tem como origem a teoria do aprendizado estatístico e sua utilização para classificação textual foi proposta por [Joachims 1998]. No contexto de classificação textual, o SVM considera cada palavra ou termo como uma dimensão em um espaço e cada documento da coleção é considerado um ponto no mesmo espaço. O algoritmo projeta planos entre os conjuntos de documentos, visando separar documentos de classes distintas. Sua versão mais simples é aplicada a problemas lineares de duas classes. Porém, existem variações que permitem a utilização do algoritmo em problemas não-lineares e multiclasses [Lupu et al. 2011] [Cortes and Vapnik 1995]. Segundo [Joachims 1998], o SVM é robusto ao *overfitting*, não necessita seleção de atributos e apresenta bom desempenho ao lidar com bases de dados de alta dimensionalidade. Essas características tornam o SVM vantajoso na classificação textual, uma vez que problemas de categorização de textos possuem alta dimensionalidade e a maioria são linearmente separáveis [Joachims 1998].

Na *Etapa de Avaliação (5)*, foi utilizada a Validação Cruzada Estratificada de 10 partições, do inglês *10-Fold Cross-Validation Stratified*, como método de validação. Esse método divide a coleção de documentos em dez partições mutuamente exclusivas e com a quantidade de exemplos de cada classe proporcional. A cada iteração, nove partições são utilizadas para treinamento e uma partição para teste. Como estimativa de erro do classificador, adotou-se a acurácia. Essa medida é calculada pela proporção de instâncias de teste que são classificadas corretamente. Outra métrica observada foi o Índice de Concordância (Índice Kappa), o qual mede o nível de concordância entre as classes verdadeiras e as classes indicadas pelo classificador. O valor de *kappa* varia entre 1 (total concordância) e valores negativos (total discordância).

5. Experimentação

Esta seção apresenta os experimentos realizados em função do processo de KDD modelado e apresenta os valores encontrados em cada uma de suas etapas. Foram testadas diferentes técnicas de atribuição de pesos e diferentes parametrizações dos algoritmos visando identificar os classificadores de melhor desempenho. Por fim, os resultados dos

melhores classificadores são avaliados e discutidos.

5.1. Coleta e Pré-Processamento

No presente trabalho, foram utilizados documentos disponíveis no site de busca do USPTO[4]. A coleta foi efetuada em outubro de 2012 utilizando-se um *web crawler* desenvolvido em linguagem Java, o qual efetuava a pesquisa por subclasses, *e.g*, CCL/702/19 para a subclasse 702/19. Para cada item retornado na busca, o *web crawler* identificava e coletava a subclasse obrigatória e as arbitrárias, número da patente, título, resumo, descrição e data da concessão. No total foram coletados 3.941 documentos. A Figura 3 exibe a distribuição de documentos por subclasse obrigatória de bioinformática.

Apesar dos documentos apresentarem estrutura completa e de ter sido coletada a descrição, optou-se por processar apenas os campos de título e resumo, uma vez que, segundo [Lupu et al. 2011] os campos de descrição e reivindicações não agregam representatividade durante a classificação automática e aumentam o custo computacional.

Durante a coleta, optou-se por efetuar o pré-processamento das palavras coletadas, assim já foram excluídos caracteres especiais (pontuações, espaços, traços, etc), *stopwords* (artigos e conectores) e, para realizar o *stemming*, aplicou-se o algoritmo de Porter [Porter 1997], por ser um dos algoritmos mais simples e eficientes [Baeza-Yates and Ribeiro-Neto 1999]. Cada palavra reduzida ao seu radical foi armazenada na tabela de termos com sua respectiva frequência e peso, segundo o método *tf*idf*. A soma das frequências de todas as palavras coletadas totaliza 260.945, as quais após a etapa de pré-processamento correspondem à 6.149 termos.

5.2. Classificação

Na etapa de classificação foram consideradas apenas as subclasses obrigatórias de cada documento, pois essa classificação é única, ou seja, cada patente é classificada por apenas uma subclasse obrigatória. A utilização das subclasses arbitrárias tornaria o problema multirrótulo, o que demandaria uma análise diferenciada da proposta deste trabalho.

Visando encontrar os valores mais adequados de acurácia e concordância (*kappa*), foram definidas seis coleções a partir da coleção original obtida no pré-processamento. A Tabela 2 exibe as coleções utilizadas para analisar os algoritmos de classificação em função dos métodos de peso utilizados e frequências mínimas de termos.

[4]http://patft.uspto.gov/netahtml/PTO/search-adv.htm

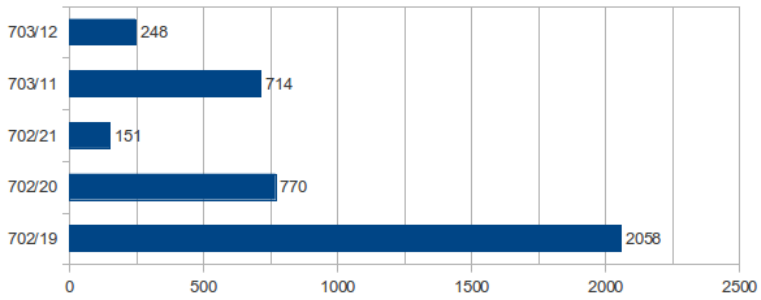

Figura 3. Distribuição de documentos por classes de bioinformática.

Coleção	Método do Peso	Frequência Mínima
tf	Frequência	1
tf-min3	Frequência	3
tf-min5	Frequência	5
tf-idf	*tf*idf*	1
tf-idf-min3	*tf*idf*	3
tf-idf-min5	*tf*idf*	5

Tabela 2. Coleções utilizadas na análise dos classificadores.

Para cada coleção foram executadas iterações com valores de parâmetros distintos, com o objetivo de identificar os classificadores e configurações de melhor desempenho. Assim, para o algoritmo NBM foi executada apenas uma iteração, pois o mesmo não dispõe de variáveis de configuração. Para o algoritmo kNN foram realizadas nove iterações, que testaram o parâmetro *k* com valores 1, 3 e 5. Esses valores foram combinados com os cálculos de distância euclidiana sem atribuição de peso e com atribuições de pesos dadas pelas fórmulas 1/distância e 1-distância. Para o algoritmo SMO foram realizadas quatro iterações testando o parâmetro *c* com os valores 0.1, 1, 10 e 100.

5.3. Resultados e Avaliação

Após as execuções dos experimentos, foram reservadas para análise no presente trabalho as iterações que apresentaram acurácia maior que 40%. A Tabela 3 exibe um comparativo dos experimentos para as seis coleções e destaca, em negrito, as maiores acurácias e índices *kappa*. Para o algoritmo NBM, o maior par de valores foi atingido na iteração da sexta coleção, denominada *tf_min5*. Nessa iteração a acurácia foi de 51,72% e a concordância igual à 0,16. Para as iterações com o kNN, os maiores valores foram encontrados na terceira iteração onde o parâmetro *k* valia 5 e não foi utilizada a atribuição de pesos para as distâncias. Nas coleções *tf_idf_min3* e *tf_min3* foram obtidos, praticamente, os mesmos valores de acurácia (45,9%) e *kappa* (0,01). Na execução do algoritmo SMO, os maiores valores foram obtidos com o parâmetro *c* valendo 0,1. As iterações nas coleções *tf_idf_min5* e *tf_min5* atingiram praticamente o mesmo par de valores, sendo 53,6% de acurácia e 0,12 de concordância.

O classificador com melhor desempenho observado no presente trabalho foi o SVM. Entretanto, na comparação entre NBM e kNN, o experimento aqui realizado obteve melhores resultados com o algoritmo NBM. Mesmo comparando valores de *F-Measure* (medida também baseada em precisão e revocação) obtidos neste trabalho, o NBM demonstrou melhor desempenho 0,65 contra 0,63 obtido com o kNN. Em [Joachims 1998], o autor compara o desempenho de algoritmos utilizados em classificação textual utilizando como medida o método *precision-recall break even point*, o qual é baseado nas medidas de precisão e revocação. Entre os algoritmos analisados estão os mesmos utilizados no presente trabalho, NBM, kNN e SVM. Os resultados do classificador proposto neste artigo corroboram o resultado de [Joachims 1998], uma vez que ele também definiu em seu estudo de caso, o SVM como o melhor classificador.

6. Conclusão

O presente trabalho utilizou um processo de KDD adaptado para modelar classificadores e investigar a classificação automática de patentes de bioinformática. Assim como observado em [Joachims 1998], o SMO apresentou melhor desempenho, mesmo em face

	NBM		(2) kNN		(3) kNN		(9) kNN		(1) SMO		(2) SMO	
Coleção	Acurácia	Kappa	Acurácia	Kappa	Acurácia	Kappa	Acurácia	Kappa	Acurácia	Kappa	Acurácia	Kappa
tf	52,01	0,05	42,53	-0,02	45,46	0	41,83	-0,01	52,08	0,1	39,58	0,01
tf_idf	41,42	0,15	42,53	-0,02	45,4	0	41,79	-0,01	52,06	0,1	39,58	0,01
tf_idf_min3	40,5	0,18	42,81	-0,01	**45,97**	**0,01**	42,09	0	53,1	0,11	40,21	0,03
tf_idf_min5	41,29	0,19	43,82	-0,03	46,78	0	41,53	-0,03	**53,66**	**0,12**	41,34	0,04
tf_min3	51,54	0,12	42,81	-0,01	**45,96**	**0,01**	42,09	0	53,1	0,11	40,22	0,03
tf_min5	**51,72**	**0,16**	43,84	-0,03	46,78	0	41,51	-0,03	**53,65**	**0,12**	41,34	0,04

(2) kNN k = 3; sem atribuição de peso para as distâncias
(3) kNN k = 5; sem atribuição de peso para as distâncias
(9) kNN k = 5; atribuição de peso = 1-distância
(1) SMO c = 0,1
(2) SMO c = 1

Tabela 3. Comparativo de desempenho das iterações com melhores valores de acurácia e correlação (*kappa*).

da alta dimensionalidade inerente ao processamento textual e sem um método específico para seleção de atributos. Entretanto, os valores de acurácia e concordância observados nos experimentos ainda se encontram abaixo do esperado.

Um possível motivo para os baixos desempenhos é a aplicação de algoritmos multiclasse. Devido à característica interdisciplinar das patentes, se faz interessante a utilização das classificações complementares, além da classificação obrigatória única em conjunto com algoritmos multirrótulo [Sousa et al. 2012].

Como trabalhos futuros, propõem-se a utilização de algoritmos multirrótulo, processamento também do campo descrição, utilização de artefatos linguísticos, como a GO (*Gene Ontology*), para a redução da dimensionalidade e uma possível equalização do número de documentos por classe. Com essas alterações, espera-se melhorar o desempenho dos classificadores testados, selecionar o melhor e integrá-lo ao sistema de busca guiada de patentes de bioinformática.

Referências

Baeza-Yates, R. and Ribeiro-Neto, B. (1999). *Modern Information Retrieval*. Addison Wesley.

Barbosa, D. B. (2003). *Uma Introdução à Propriedade Intelectual*. Lumen Juris, 2 edition.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20:273–297. 10.1007/BF00994018.

Eisinger, D., Wachter, T., Bundschus, M., Wieneke, U., and Schroeder, M. (2012). Analysis of mesh and ipc as a prerequisite for guided patent search. *Bio-Ontologies 2012*.

EPO (2011). European patent office. European Patent Office. Acesso em: Março de 2013 Disponível em: http://www.epo.org/about-us/annual-reports-statistics/annual-report/2011.html.

FAPESP (2010). Indicadores de ciência, tecnologia e inovação em sp 2010. Technical report, FAPESP. Cap. 5: Atividade de patenteamento no Brasil e no exterior; Acesso em: Março de 2013 Disponível em: http://www.fapesp.br/indicadores/2010/volume1/cap5.pdf.

INPI (2011a). Instituto nacional de propriedade industrial. Combate ao backlog de patentes é prioridade no INPI. Acesso em: Março de 2013 Disponível em: http://www.inpi.gov.br/noticias/combate-ao-backlog-de-patentes-e-prioridade-no-inpi.

INPI (2011b). Instituto nacional de propriedade industrial. Instituto Nacional de Propriedade Industrial. Acesso em: Março de 2013 Disponível em: http://www.inpi.gov.br/portal/artigo/publicacoes.

Joachims, T. (1998). Text categorization with suport vector machines: Learning with many relevant features. In *Proceedings of the 10th European Conference on Machine Learning*, ECML '98, pages 137–142, London, UK, UK. Springer-Verlag.

Lupu, M., Mayer, K., Tait, J., and Trippe, A. J. (2011). *Current Challenges in Patent Information Retrieval*. Springer.

Mukherjea, S. and Bamba, B. (2004). Biopatentminer: an information retrieval system for biomedical patents. In *VLDB '04 Proceedings of the Thirtieth international conference on Very large data bases*, volume 30, pages 1066 – 1077. Very Large Data Bases (VLDB) Endowment.

Park, H.-S. (2012). Preliminary study of bioinformatics patents and their classifications registered in the kipris database. *Genomics & Informatics*, pages 271–274.

Porter, M. F. (1997). An algorithm for suffix stripping. *Readings in information retrieval*, pages 313–316.

Reid, P. (2011). Obama's call for innovation stifled by patent office backlog. Acesso em: Março de 2013 Disponível em: http://www.cbsnews.com/8301-503544_162-20029731-503544.html?tag=mncol;lst;1.

Rodriguez, V. (2010). The backlog issue in patents: A look at the european case. *World Patent Information*, 32(4):287–290.

Sousa, F. S., Mancini, F., Teixeira, F., Araujo, G. D., Nunes, F. L. S., and Pisa, I. T. (2012). Multirrotulação automática de páginas web de saúde: uma avaliação preliminar da percepção humana. *XII Workshop de Informática Médica (WIM 2012)*.

Tan, P.-N., Steinbach, M., and Kumar, V. (2006). *Introduction to Data Mining*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.

Teixeira, F., Sousa, F. S., Araujo, G. D., Mancini, F., Araujo, L. V., and Pisa, I. T. (2012). Indexação de artigos científicos de informática em saúde por meio da competição de técnicas de extração de características. *XII Workshop de Informática Médica (WIM 2012)*.

USPTO (2005). Handbook of classification. Acesso em: Março de 2013 Disponível em: http://www.uspto.gov/web/offices/opc/documents/handbook.pdf.

USPTO (2012). Annual reports 2012. United States Trademark and Patent Office. Acesso em: Março de 2013 Disponível em: http://www.uspto.gov/about/stratplan/ar/index.jsp.

WIPO (2010a). Ipc electronic forum - project a019 - bioinformatics. Acesso em: Março de 2013 Disponível em: http://web2.wipo.int/ipc-ief/en/project/1314/A019.

WIPO (2010b). Patent applications by country of origin and by office (1995-2009). Statistical Publication. Acesso em: Março de 2013 Disponível em: www.wipo.int/export/sites/www/ipstats/en/statistics/patents.