MineraSkills: Mineração de Dados Aplicada às Vagas Anunciadas no LinkedIn Visando Definir o Perfil Profissional

Dayane Cristine M. F. Caldeira*, Ronaldo Celso Messias Correia*,
Rogério Eduardo Garcia*, Danilo Medeiros Eler*, Celso Olivete Junior*
*Departamento de Matemática e Computação
Universidade Estadual Paulista, SP, Brasil
dayanecristine.caldeira@gmail.com, {ronaldo,rogerio,daniloeler,olivete}@fct.unesp.br

***Resumo*—The content posted on the online social networks has some interesting features, such as, the wide data availability, the range of subject and constant updates, however it's important to know how to use these data to generate knowledge. This paper presents a tool that use data mining technics to explore job posts published on LinkedIn, the goal is to define the professional profile described in these job posts. The keyword extraction and the generation of association rules was employed to do that. The results allows to identify the most relevant skills and the relations between them.**

I. Introdução

A mineração de dados se tornou uma área extremamente abrangente e a prova disso é que seus métodos vem sendo aplicados em diferentes áreas com o objetivo de gerar conhecimento. Uma dessas áreas que tem se destacado é a mineração em redes sociais *online*, isso devido ao fato deste tipo de *site* gerar diariamente uma grande quantidade de dados, outro fator interessante nas redes sociais é a diversidade de conteúdo disponibilizado tendo em vista que pessoas de diferentes lugares e culturas compartilham seus conhecimentos.

Neste trabalho foram aplicadas técnicas de mineração de dados em vagas anunciadas no *site* LinkedIn, a maior rede social profissional da *web* com mais de 332 milhões de usuários com perfil cadastrado [1]. O objetivo é encontrar os requisitos que são mais solicitados nas vagas e o relacionamento entre requisitos.

É importante lembrar que aplicar a mineração de dados não é uma tarefa trivial [Fayyad et al. 1996] e o processo torna-se mais complexo quando se está lidando com dados não-estruturados, como é o caso dos dados disponibilizados pelo LinkedIn que estão na forma de texto livre, sendo assim, processos de mineração de texto e processamento de linguagem natural foram necessários para garantir a qualidade dos dados e permitir a extração de conhecimento.

Para selecionar os requisitos mais solicitados foi utilizada a extração de palavras-chave, para isso, foi implementado o algoritmo TF-IDF (*Term Frequency - Inverse Document Frequency*) que estabelece um peso para cada palavra no documento, com base no resultado gerado é possível saber a relevância das mesmas [Berry and Kogan 2010], [Hotho et al. 2005]. Já para gerar os relacionamentos entre requisitos foi empregado o clássico algoritmo para geração de regras de associação, Apriori.

Com as palavras-chave apresentadas é possível identificar mais facilmente os requisitos que estão sendo mais solicitados nas vagas analisadas, já as regras de associação encontradas podem auxiliar o usuário na tomada de decisões, pois conhecendo os relacionamentos entre palavras-chave pode-se, por exemplo, identificar requisitos que são geralmente solicitados juntos.

Este artigo está dividido da seguinte forma: a Seção II apresenta alguns trabalhos correlatos; a Seção III apresenta a ferramenta criada, MineraSkills; na seção IV é apresentado um estudo de caso; e por fim, a seção V exibe as considerações finais e trabalhos futuros.

II. Trabalhos Correlatos

O trabalho [Bradbury 2011] de Bradbury apresenta um estudo sobre as possibilidades de mineração utilizando os dados gerados pelo LinkedIn. O autor explica sobre as opções de pesquisa avançadas utilizando a interface do próprio site, além da possibilidade se trabalhar com dados não processados, seja obtendo o mesmo a partir de uma API ou os exportando em forma de arquivo CVC. São apresentadas algumas ferramentas que podem auxiliar a análise gerando diagramas e gráficos.

O artigo [Yanaze and Lopes 2014] apresenta um método para identificar e analisar as expectativas do mercado de trabalho nas áreas da engenharia elétrica e engenharia da computação, para isso foram coletadas vagas de emprego anunciadas no IEEE (*Institute of Electrical and Electronics Engineers*).

No blog oficial do LinkedIn há uma pesquisa relacionada ao ano de 2014 desenvolvida por Sohan Murthy que lista as 25 competências mais comuns nas pessoas contratadas naquele ano [4].

Em [Diaby and Viennet 2014] é proposto um sistema de recomendação de vagas de emprego para usuários das redes sociais LinkedIn e Facebook, o algoritmo proposto se baseado em um conjunto de taxonomias ao invés de utilizar o TF-IDF.

[1] https://press.linkedin.com/about-linkedin

[4] http://blog.linkedin.com/2014/12/17/the-25-hottest-skills-that-got-people-hired-in-2014/

Em [Tajbakhsh and Solouk 2014] foi implementado um sistema de recomendação de conexão baseado nas características do usuário, o LinkedIn foi utilizado como estudo de caso, o foco é voltado para a localização geográfica mas também levam em consideração outras informações entre elas as competências.

III. MINERASKILLS

A ferramenta *Mineraskills* realiza a coleta, processamento e apresentação dos dados, a mesma foi implementada utilizando a linguagem de programação Python juntamente com o framework Django. Algumas bibliotecas foram utilizadas para auxiliar a implementação, entre elas destaca-se a NLTK (Natural Language Toolkit) empregada nas tarefas de processamento de linguagem natural. Para o armazenamento dos dados foi utilizado o banco de dados não relacional MongoDB que permite armazenar dados no formato JSON.

O processamento dos dados implementado na MineraSkills pode ser dividido em quarto etapas, sendo estas, coleta de dados, pré-processamento, mineração e avaliação dos resultados. A Figura 1 apresenta uma visão geral deste processamento.

Na etapa de Coleta de Dados foi utilizada a API Job Search para obter dados a partir da rede social LinkedIn. A API, disponibilizada pelo próprio LinkedIn, retorna vagas de emprego publicadas por empresas no site. Adicionalmente, a API está fundamentada no estilo de arquitetura REST, que visa permitir a comunicação entre cliente e servidor de maneira simplificada e eficiente fazendo uso de métodos HTTP. Além disso, o site também implementa o protocolo OAuth 2.0 para garantir que somente usuários autorizados tenham acesso aos dados disponíveis no servidor [2]. Finalmente, a API retorna um arquivo XML ou JSON contendo os dados.

Uma vez que os dados são obtidos, a etapa de pré-processamento é, sem dúvida, a mais trabalhosa do processo de descoberta de conhecimento. Neste trabalho, o pré-processamenteo foi dividido nas seguintes etapas:

- Tokenização: realizada por meio da biblioteca NLTK que oferece uma função onde é possível através de expressões regulares determinar como dever ser feita a tokenização. No caso, foi determinada uma expressão regular que desconsiderasse sinais de pontuação;
- Limpeza dos Dados: realizada para retirar do texto elementos que não contribuem para a descoberta de conhecimento. Foram implementadas as seguintes etapas para eliminar os dados desnecessários: remoção de *TAG's* HTML, Remoção caracteres especiais, remoção de *stopwords*, *Part-of-Speech* (POS) *tagging* utilizando a biblioteca NLTK – trata-se de uma análise morfológica de cada *token* que determina a classe gramatical dos mesmos;
- Lematização: realizada para normalizar as palavras presentes nos textos, com o intuito de evitar variação das palavras, uma vez que o texto é curto. Para isso foi utilizado o dicionário *thesaurus* da biblioteca NLTK;
- Bigramas: adição de bigramas que apareçam no texto mais de 10 vezes, pois algumas palavras só fazem sentido se estiverem acompanhadas de outra palavra no texto. Assim, são analisados os unigramas e bigramas contidos no texto.

Após o pré-processamento, a MineraSkills realiza a mineração de dados por meio da extração de palavras-chave, com o objetivo de encontrar as palavras que melhor descrevem o conjunto de texto; e por meio da geração de regras de associação, visando encontrar relações entre as palavras mais relevantes.

Para a extração de palavras-chave o algoritmo escolhido foi o TF-IDF (*Term Frequency- Inverse Document Frequency*) que calcula a importância de uma palavra para o documento [Berry and Kogan 2010], [Hotho et al. 2005], o cálculo é realizado considerando a frequência da palavra no texto e a frequência inversa, como pode ser visto na Equação (1) abaixo.

$$tfidf_{t,d} \; = \; tf_{t,d} \times idf_t \; = \; tf_{t,d} \times log(\frac{N}{df_t}) \qquad (1)$$

Na extração de palavras-chave foi utilizado o modelo de espaço vetorial, onde o conjunto de documentos é representado na forma de matriz, em que cada linha representa um documento e as palavras são representadas nas colunas, e o valor TF-IDF é então calculado para cada palavra nos documentos. As palavras que representam os requisitos mais solicitados são aquelas que possuem o menor valor TF-IDF, tendo em vista que estas são as que aparecem em mais documentos, entretanto também foi necessário estabelecer um valor mínimo de relevância para que palavras pouco discriminatórias não sejam selecionadas. Após alguns experimentos, verificou-se que o melhor resultado é utilizando o valor 0.5.

A geração de regras de associação faz uso de dois parâmetros para determinar a relevância de uma regra gerada, o primeiro é conhecido como suporte, o mesmo pode ser definido como a porcentagem de documentos do conjunto de dados que contém a regra, o segundo é a confiança, essa medida é calculada pela razão entre o suporte e a porcentagem de documentos que contém somente um dos termos da regra [Hipp et al. 2000]. Para a extração de regras de associação a MineraSkills utiliza o algoritmo Apriori. O primeiro passo desse algoritmo consiste em calcular a frequência de cada item gerando conjuntos de itens frequentes de tamanho unitário. Os passos seguintes podem ser divididos em duas etapas, primeiramente os conjuntos gerados no passo anterior são utilizados para gerar os novos candidatos à itens frequentes, em seguida é calculado o suporte de cada candidato eliminando os que tiverem valor menor que o mínimo. Na segunda etapa é efetuada a descoberta de regras de associação, o procedimento neste caso consiste em encontrar a medida de confiança em cada conjunto de itens frequentes e se a mesma for maior do que a confiança mínima a regra é estabelecida [Agrawal and Srikant 1994].

IV. ESTUDO DE CASO

Este estudo de caso foi realizado com 7026 vagas da área de tecnologia da informação, em alguns testes as vagas foram separadas por país. Nas próximas subseções é descrito detalhes deste estudo de caso.

[2] www.oauth.net

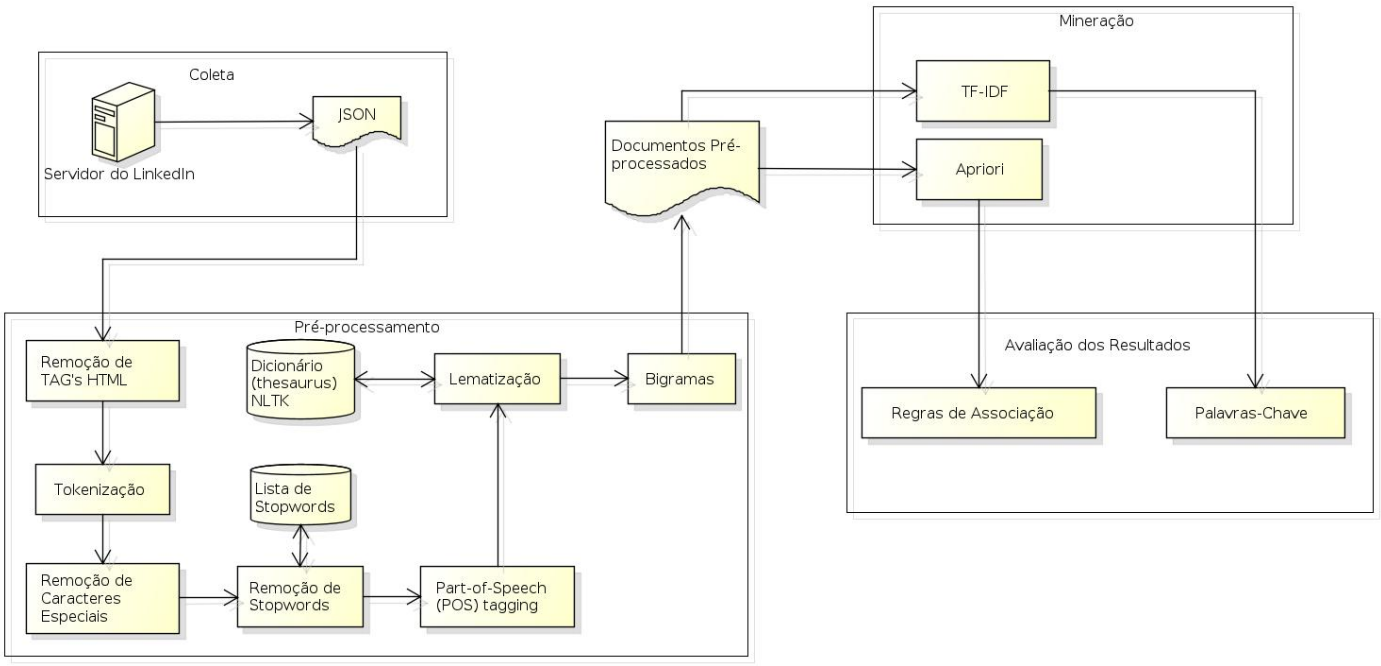


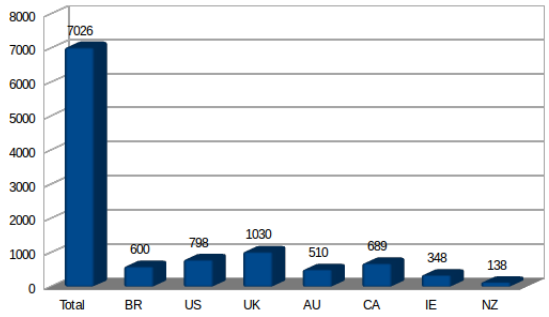
Figura 1: Modelo de Processamento de Dados Utilizado pelo Mineraskills.

A. *Filtragem e Classificação*

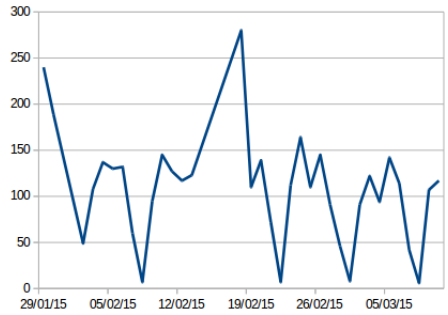
Além de obter os dados, um ponto importante do processo de coleta foi a filtragem e classificação das vagas. Ao fazer uma requisição utilizando a API do LinkedIn é possível filtrar os dados que serão coletados determinando alguns parâmetros na URL, inicialmente foram aplicados dois filtros, o primeiro por país e o segundo por palavras-chave. Os países escolhidos foram, com exceção do Brasil, locais que tem como língua oficial o inglês e possuem um número relevante de usuários e empresas cadastradas no site, sendo estes, Estados Unidos, Reino Unido, Austrália, Canadá, Irlanda, e Nova Zelândia, a finalidade deste filtro é priorizar vagas escritas em inglês e vagas para empregos no Brasil. O filtro por palavras-chave utiliza palavras relacionadas com a área de TI, entretanto este filtro falha várias vezes permitindo que documentos fora do escopo estabelecido sejam coletados. Para resolver este problema foi necessário utilizar o algoritmo de classificação Näive Bayes para garantir que somente vagas de TI fossem analisadas. O algoritmo teve um bom desempenho, conseguindo classificar corretamente 86% das vagas utilizadas como teste.

B. *Dados coletados*

Os gráficos abaixo mostram respectivamente a quantidade de vagas da área de TI (Tecnologia da Informação) escritas em inglês ou em português que foram coletadas, inicialmente as vagas foram obtidas sem o filtro por país somando um total de 2913 vagas iniciais. O segundo gráfico mostra a quantidade de novas vagas obtidas entre 29 de janeiro de 2015 e 10 de março de 2015. Para evitar sobrecarga no servidor, o LinkedIn estabelece um limite de 300 requisições por dia, sendo possível obter até 20 documentos em cada requisição. É importante ressaltar que um mesmo documento pode vir em diferentes requisições o que acaba atrapalhando a aquisição de novos documentos.



(a)



(b)

Figura 2: Quantidade de Vagas Coletadas (a) e Quantidade de Vagas Coletadas por Dia (b).

C. *Pré-processamento*

Neste estudo de caso foram aplicadas todas as tarefas de pré-processamento citadas anteriormente, ao término desta etapa os dados desnecessários já foram retirados e os demais

estão normalizados. Um efeito significativo que pode ser observado nos documentos analisados é a redução do léxico que diminui em média 92%.

Abaixo está um exemplo de texto antes do pré-processamento seguido do mesmo após a aplicação da etapa:

" Job Requirements UI Design Prototyping HTML and CSS skills Knowledge of Javascript (and jQuery in particular) for rapid prototyping and production code Strong, clear visual design skills and the ability to work with our creative team Mad communication skillz - verbal, written, and presentation Ability to work directly with Designers, Marketers, Engineers, and the rest of our team Adaptability - Our focus can sometimes change quickly Bachelor degree in relevant field; graduate degree a plus
 GrubHub Inc. is an EQUAL EMPLOYMENT OPPORTUNITY/AFFIRMATIVE ACTION employer. "

"requirement, design, prototyping, html, css, knowledge, javascript, jquery, particular, rapid, prototyping, production, code, strong, clear, visual, design, skill ability, work, creative, team, mad, communication, skillz, verbal, presentation, ability work, designer, marketer, engineer, team, adaptability, focus, degree, relevant, field, graduate, degree, grubhub, inc, equal, employment, opportunity, affirmative, action "

Pode-se constatar que foram retiradas todas as *TAG'S* HTML, caracteres especiais e *stopwords*, além disso, foi realizada a normalização das palavras restantes.

D. Algoritmo TF-IDF

Os gráficos abaixo são referentes à aplicação do algoritmo TF-IDF, foram utilizadas primeiramente todas as vagas em inglês coletadas, para em seguida separar as vagas por país, entre os sete país de onde as vagas foram coletadas três obtiveram resultados mais significativos, sendo estes, Estados Unidos, Reino Unido e Brasil,conforme pode ser visto nas imagens abaixo.

Na execução referente ao gráfico abaixo Figura 3 foi utilizado um conjunto de vagas formado a partir da junção de todas as vagas em inglês presentes no banco de dados do projeto. Foram selecionadas as vinte primeiras palavras com menor valor TF-IDF. A palavra "php5" foi a que apareceu em mais vagas, ou seja, foi o requisito mais solicitado no conjunto de vagas analisado tendo por consequência o menor valor TF-IDF, seguida pela palavra "mvc", também é possível observar um valor TF-IDF menor nas três palavras seguintes, sendo estas, "html", "sharepoint" e "sql server". As demais palavras seguem com valores considerados baixos com base nos experimentos realizados, com exceção das três últimas palavras ("social medium", "joomla", "sap").

O gráfico apresentado na imagem Figura 4 mostra o resultado da execução do algoritmo TF-IDF utilizando as vagas nos Estados Unidos, foram selecionadas as dez palavras com menor valor TF-IDF, a palavra que obteve o maior destaque TF-IDF foi "cisco" , entretanto todas as palavras, com exceção das duas últimas ("joomla" e "graphic"), obtiveram valores baixos e muito semelhantes entre si.

No caso da execução utilizando as vagas no Reino Unido, as palavras "angularjs" e "javascript" geraram os menores valores TF-IDF, sendo portanto consideradas mais relevantes, entretanto pode-se observar que as demais palavras também obtiveram valores TF-IDF considerados baixos.

O resultado gerado a partir das vagas no Brasil destacam as palavras "ensino superior" e "ruby", as quatro palavras seguintes ("soap","delphi", "sólida experiência" e "erp") obtiveram valores semelhantes, as demais também geraram valores baixos, o que indica as dez palavras apresentadas apareceram em muitas vagas do conjunto de vagas utilizado.

E. Algoritmo Apriori

Para aplicar o algoritmo Apriori foram necessárias as palavras-chave encontradas na execução do algoritmo TF-IDF utilizando as vagas em inglês, as palavras selecionadas foram utilizadas para separar os conjuntos de vagas onde o algoritmo Apriori seria aplicado, sendo que cada conjunto foi formado pelas vagas que continham a palavra-chave em questão, para auxiliar a implementação foi utilizada biblioteca Orange[4].

Na tabela Tabela I abaixo estão algumas regras que foram geradas após a execução do algoritmo, diversas outras foram encontradas, entretanto não apresentavam nenhuma informação significativa, portanto foram descartadas.

A interpretação das regras deve ser feita da seguinte maneira, o suporte indica a porcentagem de documentos em que a regra é válida e a confiança é a relação entre o suporte da regra pela porcentagem de documentos que contém o primeiro termo mas não o segundos, por exemplo, considerando a regra "ecommerce" $\rightarrow$ "javascript" o que concluímos com a regra é que 54% das vagas envolvendo e-commerce tem como requisito conhecimento em javascript, a confiança desta regra é de 100%, ou seja, todas as vezes em que palavra "javascript" apareceu também estava presente a palavra "ecommerce".

A descoberta deste tipo de relacionamento permite, por exemplo, a sugestão de competências à usuários baseada uma que o mesmo já possua. O valor de suporte mínimo passado como parâmetro para o algoritmo foi 0.4 e a confiança 0.5, estes valores foram utilizados para todas as execuções.

V. Conclusão e Trabalhos Futuros

Neste artigo foi apresentada a ferramenta MineraSkills que aplica técnicas de mineração de dados nas vagas de emprego anunciadas no LinkedIn. O fato dos dados serem coletados a partir de uma rede social online valoriza os resultados gerados, pois são dados atualizados constantemente o que garante que o cenário atual do mercado de trabalho está sendo analisado.

A abordagem de mineração de dados aplicada alcançou resultados em termo de geração de informação, mostrando palavras-chaves originadas a partir das vagas de TI anunciadas no site, assim é possível identificar áreas, ferramentas, *frameworks* e requisitos que estão se destacando. Com a divisão por país essa análise se torna mais objetiva e provavelmente mais relevante.

[4] www.orange.biolab.si

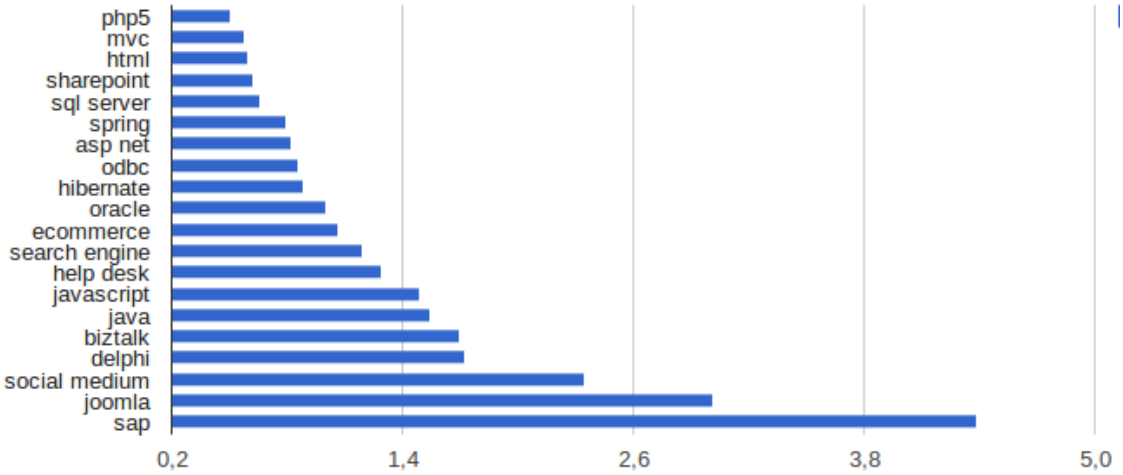


Figura 3: Palavras-chave em inglês.

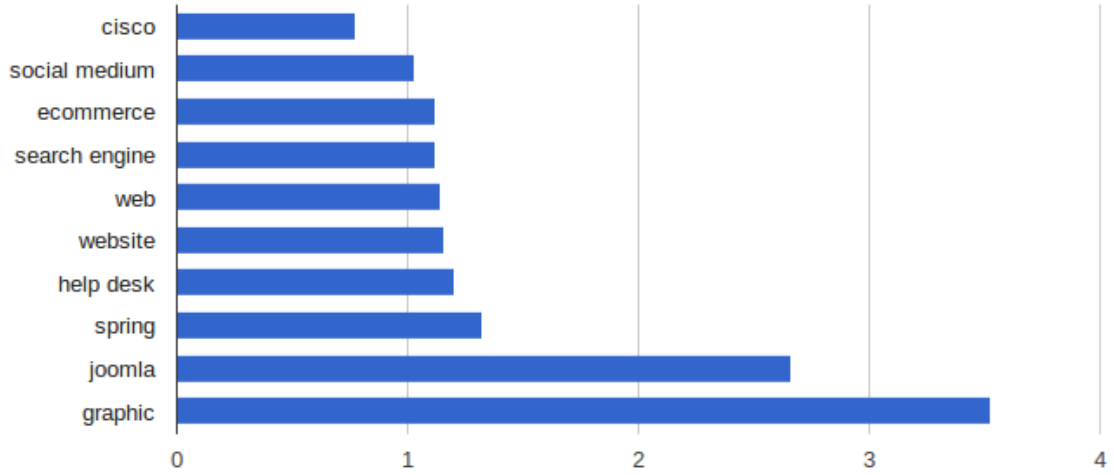


Figura 4: Palavras-chave nos Estados Unidos.

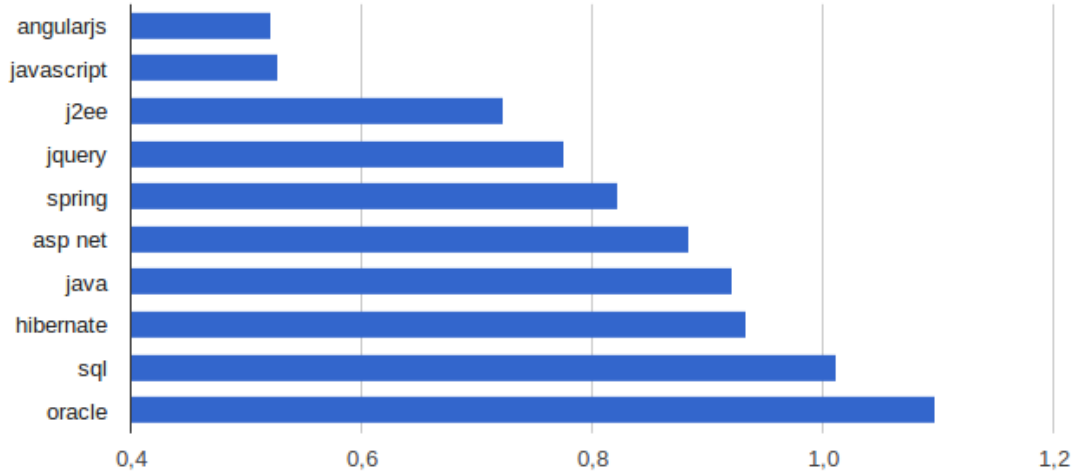


Figura 5: Palavras-chave no Reino Unido.

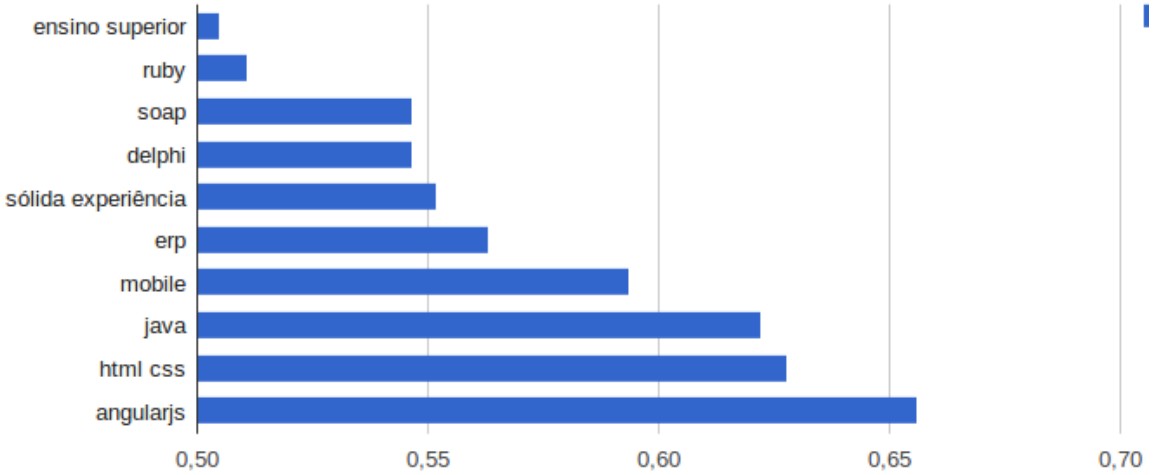


Figura 6: Palavras-chave no Brasil.

Tabela I: Regras de Associação Geradas pela Ferramenta MineraSkills

Suporte	**Confiança**	**Regra**
0.40	1.0	database → sap
0.44	1.0	sql → sap
0.48	1.0	management → sap
0.48	1.0	business → sap
0.50	0.50	joomla → analysis
0.50	0.50	joomla → javascript
0.50	0.50	joomla → web
0.56	1.0	web design → social medium
0.61	1.0	web → social medium
0.67	0.67	social medium → design
0.52	0.80	web development → javascript
0.60	1.0	sql → dba
0.52	0.52	dba → management
0.55	0.55	delphi → sql
0.50	1.0	web development → biztalk
0.67	0.67	biztalk → net development
0.75	0.75	biztalk → sql server
0.53	0.53	java → web
0.64	0.64	help desk → management
0.71	0.71	search engine → optimization
0.59	0.59	search engine → development web
0.65	0.65	search engine → javascript
0.54	1.0	ecommerce → javascript

Com a aplicação do método de geração de regras de associação foi possível identificar relações entre termos presentes nos documentos, se fez necessário implementar uma abordagem utilizando as palavras-chave dos documentos para que os resultados fossem mais satisfatório, isso devido ao fato da área de TI ser muito diversificada, o que produz textos muito diferentes entre si prejudicando a geração de regras, uma outra abordagem seria primeiramente aplicar o método de agrupamento para depois empregar o de geração de regras de associação.

Como trabalho futuro, além do agrupamento citado no paragrafo anterior, seria interessante coletar a partir de outras fontes além do LinkedIn. A extensão da área profissional para outras áreas, tais como, saúde, administração, finanças, vendas, recursos humanos etc, também é uma opção. Outro melhoria que pode ser feita é a divisão dos conjuntos de vagas por cidade visando definir o perfil profissional em cada cidade ou região.

REFERÊNCIAS

Agrawal and Srikant 1994 Agrawal, R. and Srikant, R. (1994). Fast algorithms for mining association rules in large databases. In *Proceedings of the 20th International Conference on Very Large Data Bases*. Morgan Kaufmann Publishers Inc.

Berry and Kogan 2010 Berry, M. W. and Kogan, J. (2010). *Text mining: Applications and Theory*. Chichester, U.K. Wiley.

Bradbury 2011 Bradbury, D. (2011). Data mining with linkedin.

Chiara 2003 Chiara, R. (2003). Aplicação de técnicas de data mining em logs de servidores web. Master's thesis, Universidade de São Paulo.

Diaby and Viennet 2014 Diaby, M. and Viennet, E. (2014). Taxonomy-based job recommender systems on facebook and linkedin profiles. In *Research Challenges in Information Science (RCIS)*. IEEE Computer Society.

Fayyad et al. 1996 Fayyad, U., Piatetsky-shapiro, G., Smyth, P., and Widener, T. (1996). The kdd process for extracting useful knowledge from volumes of data. *Communications of the ACM*.

Han and Kamber 2000 Han, J. and Kamber, M. (2000). *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers Inc.

Hipp et al. 2000 Hipp, J., Güntzer., U., and Nakhaeizadeh, G. (2000). Algorithms for association rule mining — a general survey and comparison.

Hotho et al. 2005 Hotho, A., Nurnberger, A., and Paass, G. (2005). A brief survey of text mining. *LDV Forum - GLDV Journal for Computational Linguistics and Language Technology*.

Keretna et al. 2013 Keretna, S., Hossny, A., and Creighton, D. (2013). Recognising user identity in twitter social networks via text mining. In *Proceedings of the 2013 IEEE International Conference on Systems, Man, and Cybernetics*. IEEE Computer Society.

Mariscal et al. 2010 Mariscal, G., Marbán, O., and Fernández, C. (2010). A survey of data mining and knowledge discovery process models and methodologies. *The Knowledge Engineering Review*.

Mislove et al. 2007 Mislove, A., Marcon, M., Gummadi, K. P., Druschel, P., and Bhattacharjee, B. (2007). Measurement and analysis of online social networks. In *In Proceedings of the 5th ACM/USENIX Internet Measurement Conference (IMC 07)*. ACM.

Sharma et al. 2012 Sharma, N., Ghosh, S., Benevenuto, F., Ganguly, N., and Gummadi, K. P. (2012). Inferring who-is-who in the twitter social network. In *Proceedings of the 2012 ACM Workshop on Workshop on Online Social Networks*. ACM.

Tajbakhsh and Solouk 2014 Tajbakhsh, M. S. and Solouk, V. (2014). Semantic geolocation friend recommendation system: Linkedin user case. In *Information and Knowledge Technology (IKT)*. IEEE Computer Society.

Yanaze and Lopes 2014 Yanaze, L. K. H. and Lopes, R. D. (2014). Transversal competencies of electrical and computing engineers considering market demand.