Uma ferramenta para semi-automatizar revisões bibliográficas

Denise E. F. Brito

Departamento de Ciência da Computação – Instituto de Ciências Exatas
Universidade Federal de Minas Gerais

denise.brit@dcc.ufmg.br

Abstract. *This work presents a framework for semi-automation of searches for scientific papers. The purpose of this framework is to provide support to researchers in the process of literature reviewing, specially in situations where the bibliographical review must follow a systematic approach. This framework allows for the user to provide a query and the desired databases where she wants to search, generating a table containing the metadata of the works returned. The queries may be done in three distinct databases: DBLP, PubMed and Cochrane.*

Resumo. *Este trabalho apresenta uma ferramenta para semi-automatização de buscas por trabalhos científicos. O intuito dessa ferramenta é de fornecer suporte a pesquisadores no processo de revisão da literatura, em especial, nas situações em que a revisão bibliográfica deve ser sistemática. A ferramenta permite que o usuário forneça uma busca e as bases de dados em que ele deseja executá-la, gerando uma tabela com os metadados dos trabalhos retornados. As buscas podem ser feitas em três bases distintas: DBLP, PubMed e Cochrane.*

1. Introdução

O processo de realizar revisões bibliográficas é fundamental em qualquer trabalho científico, pois compreende a pesquisa dos trabalhos já existentes dentro do escopo desejado. A revisão bibliográfica pode ter finalidades diferentes, dependendo do trabalho em que ela se insere. Dentre as finalidades possíveis, estão: comprovar a originalidade do trabalho que a revisão suporta, sumarizar os resultados obtidos em trabalhos anteriores sobre um determinado tema ou uma questão para estabelecer um consenso, e também fazer uma meta-análise dos estudos feitos sobre determinado assunto a fim de apontar inconsistências ou a necessidade de mais pesquisa.

A metodologia para revisão da literatura consiste, em linhas gerais, dos seguintes pontos[1]:

- Definição do escopo
- Definição dos termos de busca
- Busca em diversas fontes e bases de dados que possuam trabalhos em periódicos, livros, teses e dissertações, artigos de conferências e outros
- Refinamento por título
- Refinamento por *abstract*
- Refinamento pelo conteúdo do texto

[1] http://handbook.cochrane.org/

Em certas situações, é desejável que essa metodologia seja mais rigorosa e também que seja reprodutível, como para pesquisas na área médica, pelo seu impacto causado na sociedade, ou para a produção de *surveys*. *Surveys* são trabalhos científicos cuja contribuição consiste em compilar os principais trabalhos anteriores sobre um determinado tema e também de mostrar o que há de mais avançado (o estado-da-arte), a fim de servir de guia para outros pesquisadores em seus trabalhos futuros. Nesses casos, a revisão bibliográfica é chamada de revisão sistemática da literatura.

A revisão sistemática deve garantir uma extensa cobertura de publicações e também garantir reprodutibilidade. Além disso, é desejável que os trabalhos sejam verificados por mais de um pesquisador. A revisão sistemática, feita manualmente, demanda muito tempo e esforço dos pesquisadores. As etapas de definição do escopo, palavras-chave e de busca costumam ser iterativas, visto que se deve chegar a um ajuste fino no número de trabalhos retornados para que a pesquisa não fique incompleta, e ao mesmo tempo, não se torne inviável devido ao volume de trabalhos a serem analisados.

As bases de dados digitais facilitaram enormemente o trabalho dos pesquisadores, que antes de sua criação, dependiam das versões impressas em bibliotecas e anais. Ainda que os meios impressos não tenham sido completamente substituídos, a abrangência de bibliotecas digitais presentes na Web, principalmente em relação a trabalhos mais recentes, permite que os pesquisadores baseiem seus trabalhos, em grande parte, nos resultados obtidos dessas fontes. Algumas editoras chegam a disponibilizar uma versão *online* antes mesmo da versão impressa, que é o caso da revista Science e a versão *online* Science Express[2]. Como exemplos de bases de dados de trabalhos científicos disponíveis *online*, tem-se a Springer Link[3], com mais de 9 milhões de trabalhos de diversas áreas, publicados pela Springer; MEDLINE, via o mecanismo de busca PubMed[4], com mais de 5 mil *journals* relacionados à literatura biomédica em 2014; Cochrane Library[5], que é uma coleção de 6 bases de dados relacionadas à área de saúde; EmBASE[6], com mais de 29 milhões de registros da literatura biomédica; DBLP[7], com mais de 3 milhões de registros da área de computação.

Existem ainda ferramentas de busca como o Google Scholar[8], que retornam resultados bastante variados e que embutem estratégias próprias de ordenação dos resultados retornados pela sua relevância. No entanto, essas ferramentas são usadas com cautela por parte dos pesquisadores durante revisões sistemáticas, pois elas retornam resultados bastante heterogêneos e porque seu algoritmo de *ranking* de resultados pode ser alterado.

O objetivo deste trabalho é criar uma ferramenta para semi-automatizar o processo de busca em revisões da literatura, especialmente em revisões sistemáticas. A ferramenta criada permite que o usuário forneça os termos de busca desejados e a busca pode ser executada em três bases distintas: DBLP, MEDLINE e Cochrane Library. Os metadados disponíveis dos resultados das buscas são exportados automaticamente para uma planilha, para que os pesquisadores não tenham de visitar os *sites* de cada base de dados e exportar os resultados manualmente. A planilha gerada pode ser posteriormente compartilhada

[2] http://www.sciencemag.org/site/feature/express/introduction.xhtml#link1
[3] http://link.springer.com/
[4] http://www.ncbi.nlm.nih.gov/pubmed
[5] http://www.cochranelibrary.com
[6] http://www.elsevier.com/solutions/embase
[7] http://dblp.uni-trier.de/
[8] http://scholar.google.com

entre os participantes da revisão, para a etapa de filtragem por título. Além disso, a ferramenta provê uma análise dos resultados obtidos, em relação aos tipos de trabalhos retornados, ano de publicação e termos mais frequentes.

2. Trabalhos Relacionados

O grande volume de dados disponível nas bases de dados bibliográficas *online* despertou o interesse de vários pesquisadores de subáreas diversas da Computação. Trabalhos utilizando dados bibliográficos extraídos de bases disponíveis *online* variam da construção de ferramentas para gerenciamento de revisões sistemáticas[9] até a ferramentas de recomendação de citações [Nascimento et al. 2011], passando por extratores automáticos de referências em artigos [Alves et al. 2012] e muitas outras possibilidades.

A ferramenta StArt[10] foi criada com o objetivo de auxiliar pesquisadores durante o processo de revisões sistemáticas. Como o foco da ferramenta está no gerenciamento da revisão sistemática, fica a cargo do usuário a construção da url de busca. No entanto, com bases de dados predefinidas, essa tarefa é facilmente automatizável.

Sturm et al. criaram uma ferramenta de meta-busca chamada LitSonar [Sturm et al. 2015] para que os pesquisadores possam submeter termos de pesquisa a várias bases de dados distintas, de forma a evitar o trabalho de buscar em cada uma delas separadamente. Os autores argumentam que ferramentas de meta-busca são preferíveis a ferramentas como o Google Scholar porque estas possuem seu próprio algoritmo de *ranking*, que pode ser modificado ao longo do tempo e não é documentado, podendo levar a revisões bibliográficas que não sejam reprodutíveis. Diferentemente da StArt Tool, a ferramenta LitSonar constrói automaticamente as strings de busca para cada base de dados, a partir das palavras-chave fornecidas pelos usuários. Os dados são coletados através das *Application Programming Interfaces* (APIs) oficiais, e quando não é possível automatizar o processo, a ferramenta fornece as instruções para que o usuário colete os dados manualmente. Os autores afirmam que por questões de eficiência e por ainda ser um protótipo, o LitSonar ainda não possui a opção de exportação dos metadados dos artigos em *batch*, apenas página por página, onde uma página possui 100 registros.

Nascimento et al. propõe um *framework* [Nascimento et al. 2011] que realiza a geração automática de *queries* a partir de um artigo de entrada. Em seguida, as *queries* são utilizadas para pesquisar trabalhos semelhantes aos artigo de origem. A pesquisa é feita em bases de dados bibliográficas distintas, os resultados são deduplicados e recomendados ao usuário de acordo com a similaridade com o artigo original. Os autores utilizam apenas metadados dos artigos, pois são disponíveis publicamente mesmo quando o artigo completo não é de livre acesso. O *framework* foi proposto para o cenário em que o pesquisador possui um trabalho principal e ele deseja obter outros trabalhos que se pareçam com ele. Além disso, no cenário pensado, o usuário deseja obter uma fração de todos os resultados possíveis, por isso a etapa de recomendação.

O presente trabalho almeja auxiliar pesquisadores a realizarem revisões sistemáticas assim como a StArt Tool, no entanto, o foco está na semi-automatização das buscas, e não no gerenciamento do processo. Assim, as funcionalidades de gerenciamento serão deixadas para trabalhos futuros. Foi tomada a decisão de permitir que o usuário forneça *queries*,

[9] http://lapes.dc.ufscar.br/tools/start_tool
[10] http://lapes.dc.ufscar.br/tools/start_tool

assim como na ferramenta LitSonar e diferentemente do trabalho de [Nascimento et al. 2011], porque o cenário de utilização da ferramenta é distinto. Este trabalho é voltado para pesquisadores que já possuem as palavras-chave do seu assunto de pesquisa, e portanto, as *queries* são deixadas a cargo do especialista. Faz parte da automatização da ferramenta a construção da string de pesquisa referente a cada base de dados, a partir das *queries* fornecidas.

Uma das diferenças desse trabalho em relação ao trabalho de [Sturm et al. 2015] é que a ferramenta LitSonar coleta apenas de bases de dados que possuam uma API oficial. A ferramenta que será apresentada a seguir coleta de bases de dados que não possuem API, fazendo o *parsing* e a extração de dados da página Web, e em teoria, qualquer base de dados estruturada disponível *online* pode ser coletada com bons resultados, se o *parsing* for feito adequadamente. Outra diferença entre este trabalho e o de [Sturm et al. 2015] é que a ferramenta proposta neste trabalho permite ao usuário exportar os metadados de todos os resultados de uma só vez, em formato de planilha, e gera automaticamente uma análise dos resultados obtidos. A análise feita compreende os tipos de trabalhos obtidos e o ano de publicação, que são informações importantes para o pesquisador avaliar se o assunto está bem difundido na comunidade ou se ainda é pouco explorado, e também os termos mais frequentes no resultado da pesquisa, que podem exibir palavras-chave relacionadas para que o pesquisador retroalimente suas buscas.

3. Metodologia

A metodologia deste trabalho pode ser dividida em três etapas principais: realização da busca, de acordo com os termos passados pelo usuário, nas bases desejadas; deduplicação de trabalhos obtidos de bases diferentes, e análise dos resultados obtidos.

3.1. Busca

Para a realização da busca, o usuário deve fornecer uma string, que será interpretada e tratada para servir de parâmetro de busca nas bases de dados especificadas. A string fornecida deve seguir uma sintaxe fixa, para garantir que o termo de busca possua a mesma semântica em todas as bases. A busca pelo termo de pesquisa será feita em todos os campos disponíveis. A quantidade de resultados exportada foi limitada a 10 mil trabalhos por base de dados escolhida. Quando a busca retorna mais do que isso, é exibida uma mensagem de aviso para o usuário, com o total de trabalhos existentes na base que casam com a busca feita. A intenção dessa mensagem é de alertar o usuário para um possível refinamento da busca, pois ao seguir o processo de revisão sistemática, cada título retornado deve ser lido e considerado pelo especialista. Se a busca é muito abrangente, a revisão resultante pode se tornar imprecisa e inviável, além de privilegiar os trabalhos retornados à frente de outros por qualquer critério de ordenação. Além disso, ao fixar a quantidade de resultados, a chance de haver problemas em relação a tempo de execução e uso de memória é reduzida.

A sintaxe da busca consiste em uma string composta por palavras-chave e por operadores booleanos. As palavras-chave devem ser separadas por um espaço em branco, um operador booleano (AND, OR ou NOT) e outro espaço em branco. Os operadores booleanos devem ser escritos com letras maiúsculas para diferenciá-los dos termos de busca. Para evitar problemas na transcrição da string de busca para os buscadores das bases de dados, foi tomada a decisão de não permitir aninhamento. Isso quer dizer que

a ferramenta considerará a ordem das operações sempre da esquerda para a direita. O operador NOT não deve ser inserido no início da string ou após outro operador booleano, caso contrário, o resultado é indeterminado. Todas as bases de dados farão a busca considerando as palavras-chave passadas como prefixos. A introdução de símbolos na sintaxe para representar casamento exato de strings, assim como a opção de selecionar os campos nos quais buscar pelas palavras-chave, foi deixada para trabalhos futuros.

As bases de dados escolhidas para a implementação da meta-busca foram DBLP, Cochrane Library e PubMed, que permite a busca de trabalhos presentes na MedLine e outros da área biomédica. Cada uma das bases possui peculiaridades na busca e a extração dos dados foi feita separadamente. Além disso, as sintaxes das strings de busca também são distintas.

Os campos selecionados para serem extraídos foram: título do trabalho; nome dos autores; ano de publicação; tipo da publicação, quando existente; base de dados da qual o registro foi extraído; nome do veículo de publicação; volume da publicação, quando existente; edição da publicação, quando presente, e as páginas correspondentes ao trabalho no veículo de publicação, quando informadas. O tipo da publicação pode ser "Journal Articles", "Editorship", "In Proceedings", "Informal" , "Book", "Editorial", "Letter", entre outros.

Nem sempre é possível identificar o tipo do registro, principalmente quando a base não o fornece explicitamente, ou quando a base fornece múltiplas informações separadas por pontuação. Assim, o campo do tipo de publicação é processado de forma simples, atribuindo "Journal Articles" se essa expressão aparece em qualquer parte da string de tipo; caso contrário, procura por "Article". Em terceiro lugar, procura por "Editorial", e em quarto, procura por "Letter". Se nenhum desses termos for encontrado, particiona-se a string de tipo pela pontuação e o primeiro valor é atribuído ao campo. Foi feito um esforço para manter a consistência dos tipos presentes em diferentes bases de dados. No caso em que a base de dados não provê nenhuma informação referente ao tipo do registro, o campo permanece vazio.

A escolha dos campos foi motivada puramente pela disponibilidade em todas as bases. Garante-se assim que o usuário obtenha resultados de trabalhos disponíveis e restritos. Em algumas bases, por exemplo, na DBLP, nem sempre o *abstract* está disponível. Para homogeneizar ao máximo os dados de bases distintas, foi tomada a decisão de não extrair o *abstract*. Os metadados extraídos podem ser de grande ajuda para o pesquisador, para que ele realize a segunda etapa da revisão sistemática, que consiste em filtrar os resultados da busca pelo título. Assim, a quantidade de trabalhos restantes que são restritos já estará significativamente reduzida.

DBLP Para a DBLP, a busca foi feita utilizando a opção *CompleteSearch*. A extração dos dados foi feita a partir de *scraping* da página de resultados. Na string de pesquisa, os operadores AND são substituídos por espaços, os operadores OR, pelo símbolo |, e o NOT, por um símbolo - à frente do termo. Abaixo, há um exemplo de string de pesquisa fornecida pelo usuário:

diabetes OR obesity AND child NOT coronary

essa string significa que o usuário deseja todos os trabalhos em que sejam encontrados os prefixos "diabetes" ou "obesity", e além disso, eles devem conter o prefixo "child" e não devem conter "coronary". A seguir, o resultado da conversão da string fornecida pelo usuário para o formato de busca da DBLP:

diabetes|obesity child -coronary

Cochrane Library Para a Cochrane Library, os dados foram obtidos através da exportação da própria ferramenta e posteriormente, o resultado foi desmembrado e colocado no mesmo formato que as demais bases. Por isso, foram exportados apenas os dados de citação, e não de *abstracts*. O limite de 10 mil resultados foi aplicado a cada uma das fontes de dados da Cochrane Library: Cochrane Reviews, Other Reviews, Trials, Methods Studies, Technology Assessments e Economic Evaluations. Foram utilizadas *wildcards* (símbolos de expressões regulares para aumentar a capacidade de expressão da busca) para determinar que os termos passados correspondem a prefixos e a busca por palavras derivadas foi desativada, para não causar incompatibilidade com os resultados das demais bases. Abaixo, o resultado da conversão da string de exemplo para o formato de busca da Cochrane Library:

(((diabetes*) OR obesity*) AND child*) NOT coronary*

PubMed Para a PubMed, a *National Center for Biotechnology Information* (NCBI) disponibiliza a ferramenta *Entrez* [11], que funciona como uma API. É possível submeter vários tipos de requisições, entre elas, buscar por palavras-chave e obter um sumário de um trabalho a partir do seu identificador. Os dados então foram extraídos e colocados no formato padrão utilizado na ferramenta deste trabalho. A PubMed faz a extração automática da árvore dos termos MeSH [12]. A seguir, a conversão da string de exemplo para o formato da PubMed:

(((diabetes) OR obesity) AND child) NOT coronary

3.2. Deduplicação

A etapa de deduplicação consiste em remover registros equivalentes dos resultados. Duplicatas podem acontecer apenas entre bases de dados diferentes, pois nas bases disponíveis *online*, o mesmo trabalho não é indexado mais de uma vez.

No domínio de dados bibliográficos, um registro pode ser identificado por uma tupla $R = (T, A, Y)$ onde T corresponde ao título do trabalho, A corresponde à lista de autores e Y, ao ano de publicação. Nesse caso, basta uma simples heurística para remover grande parte das duplicatas. Esses campos são comparados, e se são exatamente iguais, com os autores podendo aparecer em ordem diferente, então é mantida apenas uma cópia do registro. Foi tomada a decisão de não utilizar a distância entre strings para decidir se dois trabalhos são iguais, pois é bastante comum encontrar conjuntos de trabalhos sobre o mesmo tema, publicados pelos mesmos autores, no caso de trabalhos curtos que são estendidos posteriormente, com nomes semelhantes.

[11] http://www.ncbi.nlm.nih.gov/books/NBK25500/
[12] http://www.nlm.nih.gov/mesh/meshhome.html

É preciso ressaltar que nem todas as duplicatas serão identificadas, principalmente quando os trabalhos não são publicados em inglês. Em alguns casos, as bases de dados mantêm o nome traduzido e algum marcador indicando o idioma original entre colchetes, por exemplo, enquanto que em outras, é mantido o título original, com a codificação UTF-8 para suportar outros caracteres não presentes na língua inglesa. Além disso, a grafia dos nomes dos autores pode estar em formatos diferentes, o que seria solucionado apenas com um algoritmo de casamento de dados, que está além do escopo deste trabalho.

3.3. Análise

Após as etapas de busca e deduplicação, a ferramenta provê uma análise dos resultados obtidos, para que o pesquisador possua um *feedback* sobre a busca realizada. Essa análise considera as palavras encontradas, o ano de publicação dos trabalhos e o tipo de trabalhos encontrados.

As palavras mais frequentes encontradas nos títulos dos trabalhos podem fornecer ao pesquisador ideias de outras palavras-chave que podem ser incluídas na string de busca. Para fazer essa análise, o título dos trabalhos é passado por alguns filtros de strings. Primeiramente, caracteres como acentos gráficos e símbolos que não estão na tabela ASCII são transliterados. A seguir, são removidas todas as pontuações, e por último, as palavras mais comuns que aparecem em todos os tipos de assuntos (*stop words*). A remoção de *stop words* é feita apenas para o inglês, pois é o idioma principal das bases utilizadas para esse trabalho.

A partir da string contendo apenas palavras significativas e bem formatadas, estas são separadas por espaço e a frequência dos termos na base é calculada, sendo incrementada no máximo uma vez por registro. Um gráfico de barras é feito com os 20 termos mais frequentes. São gerados outros gráficos de barras, um para exibir a distribuição dos trabalhos ao longo dos anos, e outro para mostrar os tipos de trabalhos mais frequentes. Essas informações podem ajudar o pesquisador a inferir o quão recente é o tópico de pesquisa, sua maturidade e como está o interesse da academia sobre ele.

4. Resultados

A implementação da ferramenta foi feita utilizando Python, e a ferramenta pode ser executada com a seguinte linha de comando:

python SearchPapers.py <infile> <outfile> deduplicate

onde infile é o nome do arquivo contendo a string de busca na primeira linha, e cada linha seguinte representa uma base de dados. As bases de dados podem ser “DBLP”, “Cochrane Library” e “Pubmed”. Outfile representa o nome do arquivo de saída, sem a extensão. Ele conterá, ao final da execução, os trabalhos resultantes em formato csv, com campos separados por tabulação e codificação UTF-8, podendo ser manipulado pelo usuário através de aplicações de planilhas. A complexidade temporal da heurística de deduplicação é quadrática em relação ao número de registros obtidos, tornando a execução muito lenta em casos onde se tem muitos resultados. Assim, o usuário possui a opção de desabilitá-la, simplesmente omitindo o último parâmetro na execução.

Abaixo, encontra-se um exemplo dos resultados obtidos pelo usuário a partir da string de busca dada de exemplo na seção 3.1:

Title: Integrated and Personalized Diabetes Coach for Children. Authors: Andy Harris, Arjan Durresi, Mihran Tuceryan, Tamara S. Hannon Database: DBLP Pages: 31-35 Series: AINA Workshops Type: Conference and Workshop Papers Year: 2015

Title: The challenges of real-world implementation of web-based shared care software: the HopSCOTCH Shared-Care Obesity Trial in Children. Authors: Kate Lycett, Gary Wittert, Jane Gunn, Cathy Hutton, Susan A. Clifford, Melissa Wake Database: DBLP Pages: 61 Periodical: BMC Med. Inf. & Decision Making Publication Volume: 14 Type: Journal Articles Year: 2014

Title: A NOS3 polymorphism determines endothelial response to folate in children with type 1 diabetes or obesity. Authors: Wiltshire EJ, Pena AS, Mackenzie K, Bose-Sundernathan T, Gent R, Couper JJ Database: Cochrane Library Pages: 319-325.e1 Periodical: Journal of pediatrics Publication Issue: 2 // () *National Health and Medical Research Council* Publication Volume: 166 Type: Journal Articles Year: 2015

Title: Behavioral counseling to prevent childhood obesity–study protocol of a pragmatic trial in maternity and child health care. Authors: Mustila T, Keskinen P, Luoto R Database: Cochrane Library Pages: 93 Periodical: BMC pediatrics Publication Volume: 12 Type: Journal Articles Year: 2012

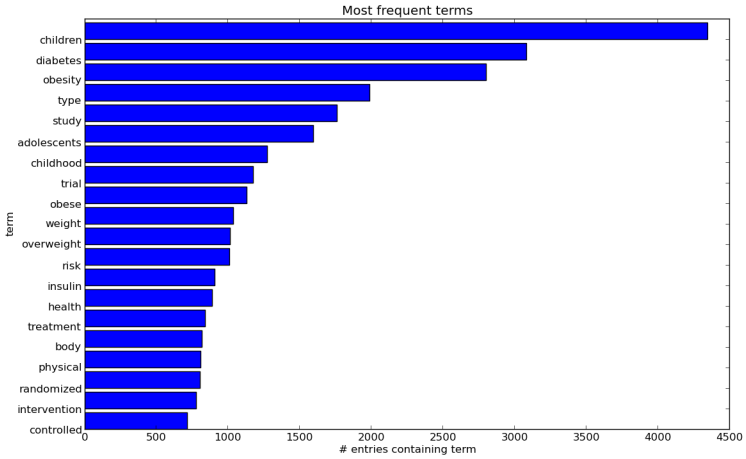
Title: Glucocorticoid-induced preterm birth and neonatal hyperglycemia alter ovine beta cell development. Authors: Bansal A, Bloomfield FH, Connor KL, Dragunow M, Thorstensen EB, Oliver MH, Sloboda DM, Harding JE, Alsweiler JM Database: PubMed Pages: en20151095 Periodical: Endocrinology Type: Journal Articles Year: 2015

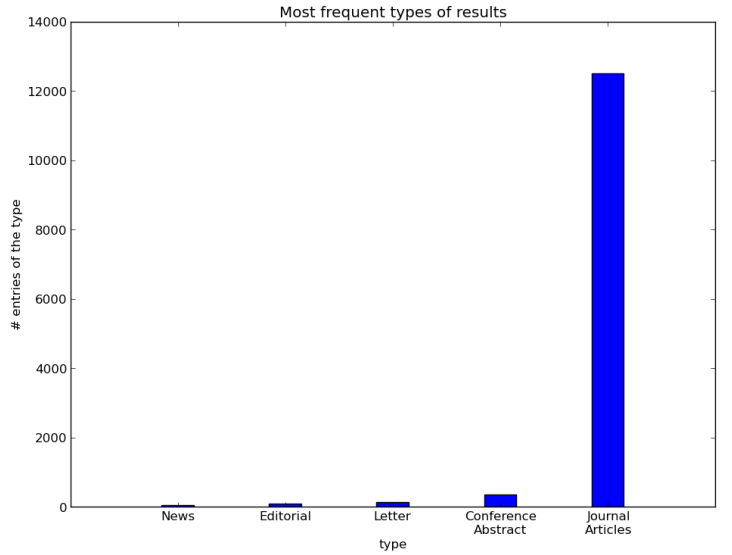
Title: Assessing Child Obesity and Physical Activity in a Hard-to-Reach Population in California's Central Valley, 2012-2013. Authors: Schaefer SE, Camacho-Gomez R, Sadeghi B, Kaiser L, German JB, de la Torre A Database: PubMed Pages: E117 Periodical: Preventing chronic disease Publication Volume: 12 Type: Journal Articles Year: 2015

A figura 1 exibe os gráficos gerados a partir do módulo de análise. Nos termos mais frequentes, aparecem *risk*, *insulin*, *intervention*, que podem ser incorporados às próximas buscas. Os gráficos dos tipos de trabalhos e da distribuição das publicações ao longo dos anos mostram que o tópico buscado é antigo, no entanto, vem despertando maior interesse da academia nos últimos anos.

Para comprovar a efetividade da análise, é interessante contrastar os resultados entre buscas diferentes. A figura 2 mostra os gráficos de análise resultantes da busca pelo termo "ebola" nas bases da Cochrane Library e PubMed. É possível notar que para essa busca, outros termos se tornam mais comuns, como "outbreak", "fever" e "viral". Além disso, a distribuição dos tipos de trabalhos se modifica, aparecendo mais publicações do tipo *News*, em comparação com a busca por obesidade e diabetes em crianças. Por fim, na distribuição dos trabalhos ao longo dos anos, emerge um comportamento de oscilação, diferentemente da busca anterior. Esse comportamento pode estar correlacionado às ocorrências da doença nas últimas décadas[13].

[13] http://www.cdc.gov/vhf/ebola/outbreaks/history/chronology.html





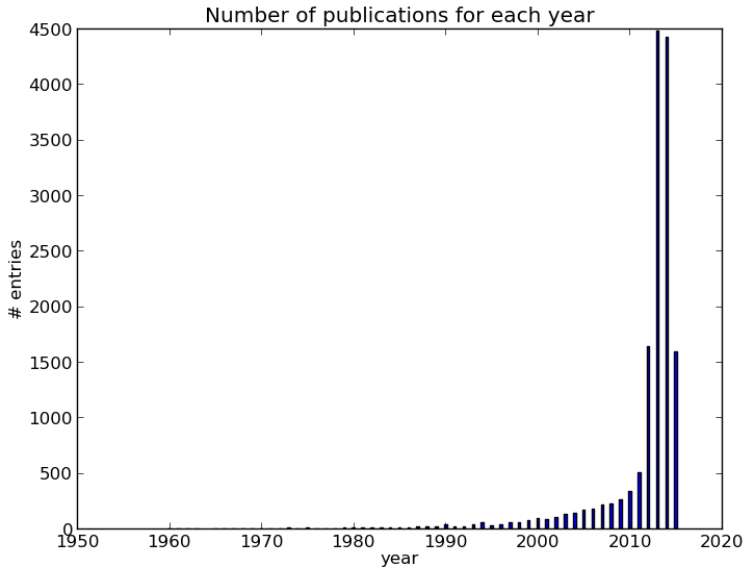
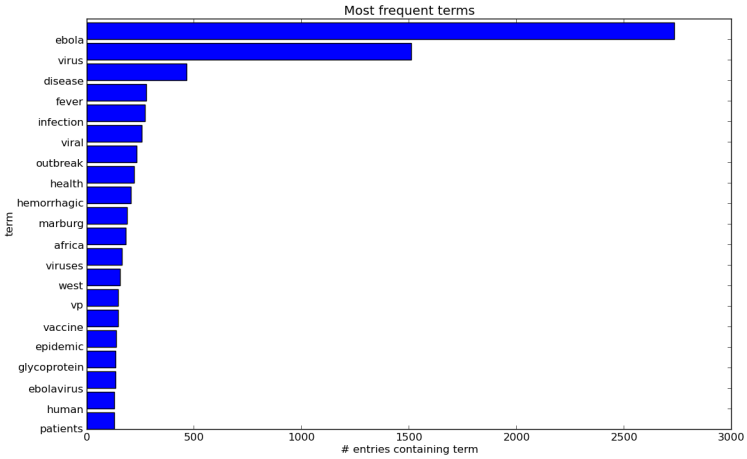
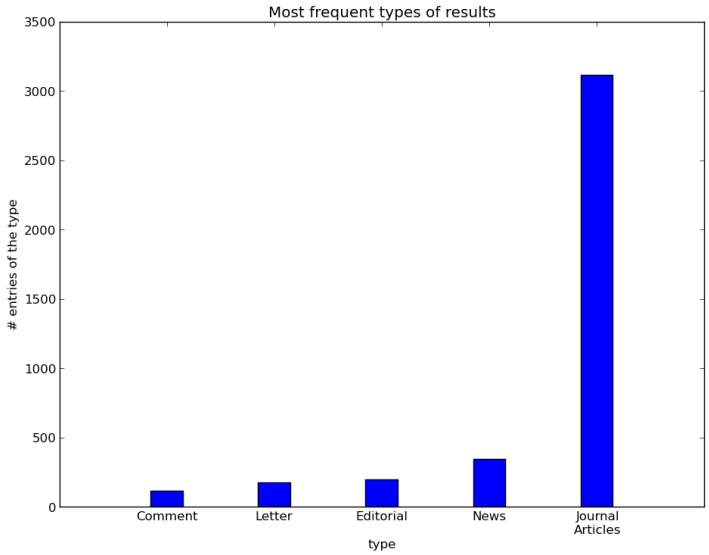


Figura 1. Resultados da análise para a string de busca do exemplo.





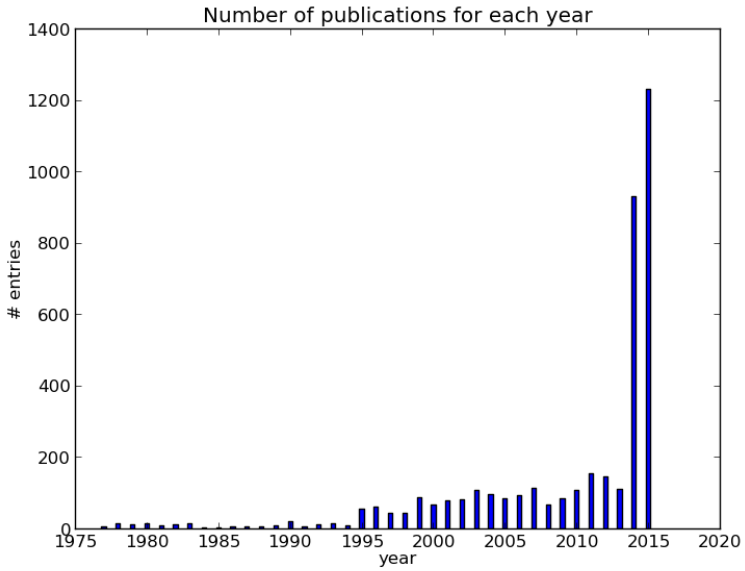


Figura 2. Resultados da análise para a busca por “ebola”.

5. Conclusões

Neste trabalho foi apresentada uma ferramenta de semi-automatização de revisões bibliográficas. Além da extração em conjunto dos metadados dos trabalhos retornados, a ferramenta provê uma análise simples, porém de grande utilidade para o usuário.

Esta é a primeira versão da ferramenta, em forma de protótipo. Existem muitos pontos que podem ser explorados em trabalhos futuros, como a criação de uma interface gráfica, a expansão da expressividade da string de busca fornecida, a inclusão de outras bases disponíveis na Web e a utilização de um algoritmo de casamento de dados mais sofisticado para realizar a deduplicação.

Apesar da simplicidade da análise realizada, os resultados mostram que é possível obter informações valiosas para o pesquisador, que podem ser utilizadas para incrementar a busca ou para prover uma noção geral sobre o tema.

Referências

Alves, N., Dueire Lins, R., and Lencastre, M. (2012). A strategy for automatically extracting references from pdf documents. In *Document Analysis Systems (DAS), 2012 10th IAPR International Workshop on*, pages 435–439.

Nascimento, C., Laender, A. H., da Silva, A. S., and Gonçalves, M. A. (2011). A source independent framework for research paper recommendation. In *Proceedings of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries*, JCDL '11, pages 297–306, New York, NY, USA. ACM.

Sturm, B., Schneider, S., and Sunyaev, A. (2015). Leave no stone unturned: Introducing a revolutionary meta-search tool for rigorous and efficient systematic literature searches. In *Proceedings of the 23rd European Conference on Information Systems (ECIS 2015)*, page 1–10, Münster, Germany.