Multirrotulação automática de páginas web de saúde: uma avaliação preliminar da percepção humana

Fernando S. Sousa[1], Felipe Mancini[2], Fabio O. Teixeira[1], Gabriela D. de Araujo[1], Fátima de L. dos S. Nunes[3], Ivan T. Pisa[1]

[1]Departamento de Informática em Saúde – Universidade Federal de São Paulo (UNIFESP) – São Paulo – SP – Brasil.

[2]Instituto Federal de Educação, Ciência e Tecnologia de São Paulo (IFSP) – Guarulhos – SP – Brasil.

[3]Escola de Artes, Ciências e Humanidades – Universidade de São Paulo (USP) – São Paulo – SP – Brasil.

buscasaude@unifesp.br

Abstract. *Lay people show difficult when they look for health information on Web. This study evaluated the adequacy of automatic multi-label suggestion for health web pages in Brazilian Portuguese language. We collected 57 health web pages and asked 21 volunteers to evaluate them. We measured the recall, consensus between evaluators and consensus between evaluators and automatic classifiers. Recall reached 100%, with high consensus between evaluators to the 5 most relevant categories, suggesting that the automatic multi-labeling of health Web pages helps information retrieval by lay people.*

Resumo. *Pessoas leigas apresentam dificuldades quando procuram por informações sobre saúde na web. Este estudo avaliou a adequação da sugestão automática de multirrótulos para páginas web de saúde em português brasileiro. Foram coletadas 57 páginas web de saúde e convidados 21 voluntários para a classificação manual. Mensurou-se a revocação, consenso entre avaliadores, e consenso entre avaliadores e classificadores automáticos Naive Bayes e Journal Descriptor Indexing. A revocação atingiu 100%, com alto consenso entre avaliadores para as 5 categorias mais relevantes, sugerindo que a multirrotulação automática de páginas web de saúde colabora com recuperação de informação por pessoas leigas.*

1. Introdução

O crescimento acelerado da web, devido principalmente à popularização de ferramentas que facilitam a criação e publicação de conteúdo [Breitman et al. 2006], fez com que a quantidade de páginas indexadas pelos principais buscadores chegasse à ordem de 48 bilhões [Kunder 2012]. Consequentemente há um aumento da dificuldade de se encontrar a informação desejada, além de dificultar a avaliação, por parte dos usuários, da relevância e confiança das informações encontradas [Fogg et al. 2003]. Parte das dificuldades é oriunda das estratégias utilizadas pelos usuários em suas buscas, quando são empregados de 2 a 3 termos, acessados poucos resultados retornados e raramente

utilizados mecanismos de buscas avançada [Toms and Latter 2007 Wolfram et al. 2009]. Quando se observa um usuário procurando por informação sobre saúde na Web, o padrão comportamental é semelhante [Zeng et al. 2002].

Procura por temas de saúde são frequentes na web. No ano de 2010, 35% das atividades de usuários da Web no Brasil estavam relacionadas à procura de informação sobre saúde [Barbosa 2011]. Já nos Estados Unidos, o relatório da Pew Research Center's Internet & American Life Project diz que 80% das pessoas procuraram por algum tipo de informação em saúde no ano de 2010, sendo doenças ou problemas de saúde e modalidades de tratamento os assuntos mais procurados [Fox 2011]. Nessas buscas, entretanto, os usuários com frequência chegam a conclusões incorretas [Keselman et al. 2008], principalmente pelo baixo nível de conhecimento dos mesmos sobre os assuntos de saúde pesquisados e pelo fato da ferramenta de busca não retornar resultados relevantes.

Outro fator que contribui com os erros são os próprios buscadores de propósito geral como o Google, que não são completamente efetivos para recuperar conteúdo relevante sobre saúde [Schembri and Schober 2009 Tang and Ng 2006], principalmente devido à grande quantidade de informações e a baixa relevância do que foi recuperado.

Portanto, o estudo e desenvolvimento de estratégias e ferramentas para classificar automaticamente conteúdos web com informações sobre saúde é importante para direcionar melhor os consumidores de informação em saúde em suas buscas na web [Stvilia et al. 2009]. Além do mais, a própria informação sobre saúde pode se dividir em diversas categorias. Para um paciente pode ser importante encontrar páginas de locais que tratam uma anomalia, classificadas, por exemplo, em uma categoria denominada "Clínicas e Hospitais". Um profissional da saúde pode estar interessado em encontrar colegas que possam auxiliá-lo em um caso difícil, cujas páginas podem estar, por exemplo, em uma categoria denominada "Profissionais".

No entanto, o conteúdo não determinístico dos textos web pode trazer características multicategóricas para uma página web, que a insira em mais de uma categoria. Por exemplo, uma página de saúde com seu conteúdo classificado como "Clínicas e Hospitais" por um diretório web ou um especialista, pode também trazer informações relevantes sobre "Profissionais". Percebe-se, adicionalmente, que usuários também costumam identificar a relação das páginas web com mais de uma categoria [Santini 2008]. Portanto, é relevante a investigação de métodos que possam sugerir uma classificação em multirrótulos (mais de uma categoria por documento) de conteúdos web de saúde com o objetivo de melhor representar a informação disponibilizada para o usuário por meio das ferramentas de busca.

O objetivo deste trabalho foi fazer uma avaliação preliminar da adequação da sugestão de vários rótulos (multirrotulação) [Qi and Davison 2009] por classificadores automáticos, comparando com a percepção humana para páginas web de saúde em português brasileiro.

O restante do documento está organizado da seguinte maneira: na seção 2 são descritos os métodos utilizados neste trabalho; na seção 3 são apresentados os resultados obtidos e é realizada a discussão dos mesmos; na seção 4 é apresentada a conclusão.

2. Métodos

Este trabalho faz parte de um projeto maior, denominado Busca Saúde [Mancini and Falcão and et al. 2010 Mancini and Sousa and et al. 2010], em desenvolvimento na UNIFESP, que tem como objetivo principal prover para o usuário leigo uma melhor experiência ao realizar buscas sobre saúde na web.

Para o presente trabalho foram coletadas 57 páginas web de saúde em português brasileiro do diretório web *Open Directory Project* (http://www.dmoz.org), representando 19 subcategorias da área de saúde. Portanto, cada categoria foi representada por 3 páginas web. Foram convidados 21 voluntários para multirrotular 20 páginas web escolhidas aleatoriamente. Estes foram instruídos a acessar um web site desenvolvido para auxiliar no experimento contendo uma lista com os links das 20 páginas web sorteadas para o voluntário. Os voluntários foram instruídos a abrir cada um dos links e avaliá-los de acordo com a sua percepção, sem nenhuma intervenção do pesquisador. Eles tinham total liberdade de escolher quantas categorias julgassem pertinentes para cada página web, bem como de mudar sua opinião quanto às escolhas.

A partir das escolhas dos voluntários foi criado um ranking de relevância de categorias [Humphrey et al. 2009 Sebastiani 2002], formado pela quantidade de seleções de cada categoria para cada página por todos os avaliadores. A categoria com mais seleções pelos avaliadores para uma página foi considerada como sua primeira categoria do ranking [Rosso 2005], e assim foi feito para cada posição do ranking, até a quinta posição.

A primeira medida de avaliação utilizada foi a revocação [Sebastiani 2002], ou seja, avaliou-se o acerto dos voluntários, frente à categoria original das páginas web selecionadas. A segunda avaliação mediu o consenso entre as categorias assinaladas pelos avaliadores e as categorias inferidas por classificadores automáticos. O consenso foi definido como a porcentagem de categorias que foram assinaladas ou inferidas tanto pela avaliação humana quanto por um dos classificadores automáticos até a quinta posição de seus respectivos rankings, independentemente da posição em que as categorias aparecem. Para comparação da classificação por voluntários com uma classificação automática, foram utilizados os resultados de trabalhos anteriores [Sousa 2011 Sousa et al. 2012], que utilizaram a mesma base de páginas web, e cujos classificadores utilizados geram como saída um ranking das categorias mais relevantes para uma página, possibilitando a sugestão de multirrótulos.

Uma terceira avaliação foi realizada para verificar a concordância entre os avaliadores, medida de acordo com o consenso entre os mesmos [Rosso 2005 Santini 2008], ou seja, a porcentagem de avaliadores que assinalaram a mesma categoria para cada página. O valor do consenso de uma página é dado pela categoria que obteve o maior consenso para a mesma.

3. Resultados e Discussão

Os participantes deste experimento eram graduados (4 voluntários) ou pós-graduados (17 voluntários). Dos 21 voluntários, 1 era da área de educação, 2 de informática, 11 de informática em saúde, 4 de odontologia, 2 de saúde e 1 de outras áreas. Com 20 páginas

sorteadas para cada um, obteve-se 21 páginas com 8 avaliações e 36 páginas com 7 avaliações.

Os resultados dos classificadores de padrões utilizados em trabalhos anteriores [Sousa 2011, Sousa et al. 2012] foram resumidos na Tabela 1. Os classificadores utilizados foram: Naive Bayes [John and Langley 1995], com extração de atributos por: ocorrência do termo (nb-to), ocorrência binária (nb-bo), frequência dos termos (nb-tf), e tf.idf (nb-tfidf) [Salton and Buckley 1988]; e Journal Descriptor Indexing [Humphrey et al. 2009], com extração de atributos por contagem de coocorrência por palavras (jdi-wc), e contagem de coocorrência por documentos (jdi-dc) [Humphrey et al. 2009].

Tabela 1. Resultados da revocação para os classificadores utilizados [Sousa 2011 Sousa et al. 2012].

	Posição do Ranking de Relevância				
Classificador	**1**	**2**	**3**	**4**	**5**
nb-tf	**0,64**	**0,75**	**0,80**	**0,88**	**0,92**
nb-tfidf	**0,74**	**0,82**	**0,88**	**0,94**	**0,96**
jdi-wc	**0,84**	**0,91**	**0,94**	**0,96**	**0,97**
jdi-dc	**0,85**	**0,92**	**0,95**	**0,96**	**0,97**
nb-bo	**0,87**	**0,91**	**0,93**	**0,95**	**0,96**
nb-to	**0,91**	**0,95**	**0,96**	**0,97**	**0,98**

Na Figura 1 é mostrado o comportamento da revocação para as 5 posições mais relevantes do ranking de relevância de categorias, tanto para a classificação humana quanto para a automática. Quando observa-se a primeira posição do ranking de relevância, ou seja, a categoria mais escolhida para as páginas pelos avaliadores e a mais relevante segundo os classificadores, o desempenho da classificação humana é bem menor que o desempenho dos classificadores automáticos, atingindo apenas 0,51. Porém, aumentando a tolerância ao medir a revocação na quarta categoria mais relevante o desempenho da classificação pelos avaliadores é melhor que qualquer um dos classificadores automáticos utilizados, chegando a 100% de acerto. Dessa forma, há indícios de que, no conjunto utilizado, apenas um rótulo para as páginas web de saúde em português brasileiro não é o suficiente para satisfazer as expectativas dos usuários, sendo necessária uma multirrotulação para que esta expectativa seja satisfeita [Santini 2008].

Outra comparação realizada foi a medida de consenso entre avaliadores e classificadores automáticos. Aqui o consenso foi definido como a porcentagem de categorias que foram assinaladas ou inferidas tanto pelos avaliadores quanto por um dos classificadores automáticos até a quinta categoria mais relevante, independentemente da posição em que as categorias aparecem.

A Figura 2 ilustra o histograma da quantidade de páginas para cada valor de consenso entre os classificadores automáticos e os avaliadores. Como se considerou as cinco categorias mais relevantes dos rankings, cada página poderia apresentar um consenso de 0%, 20%, 40%, 60%, 80% e 100%, representando respectivamente 0, 1, 2, 3, 4 e 5 categorias de consenso entre os voluntários e um classificador. Para cinco dos seis classificadores automáticos utilizados neste trabalho, a maioria das páginas obteve

um consenso de 40%, ou seja, duas categorias são assinaladas tanto pelos avaliadores quanto pelos classificadores. A exceção foi o classificador jdi-dc, cujo consenso foi de 60% para a maioria das páginas (33 de 57 páginas). Além disso, este foi o único classificador que conseguiu um consenso de 100% em uma das páginas e também atingiu um resultado bastante superior ao dos outros classificadores para o consenso de 80%, conseguindo uma média de 62,11%. Portanto, é provável que o classificador jdi-dc consiga mapear melhor a percepção que consumidores de saúde têm sobre as categorias de páginas web de saúde em português-brasileiro.

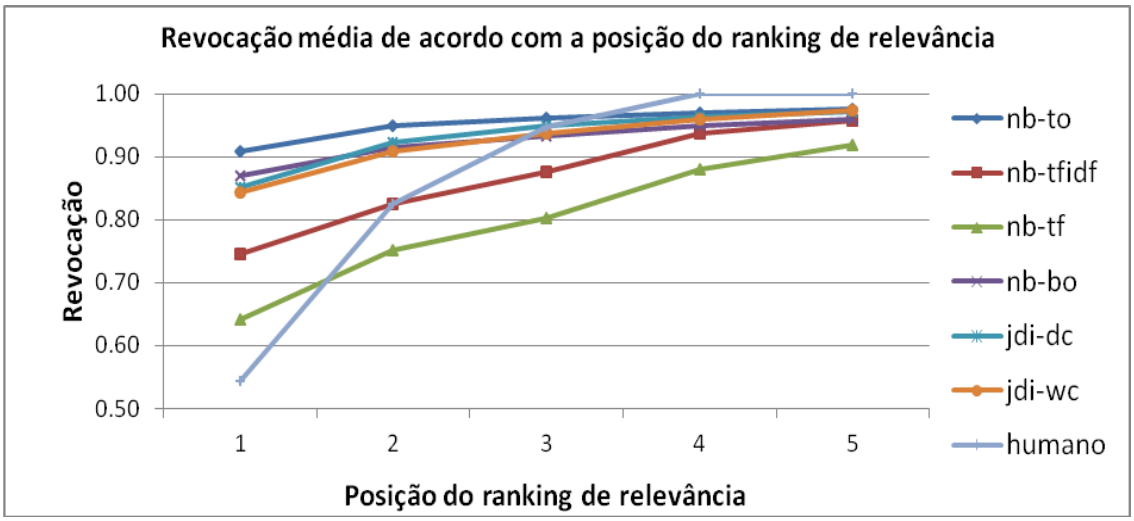

Figura 1. Comparação entre a revocação dos classificadores e da classificação humana para as cinco primeiras posições do ranking de relevância.

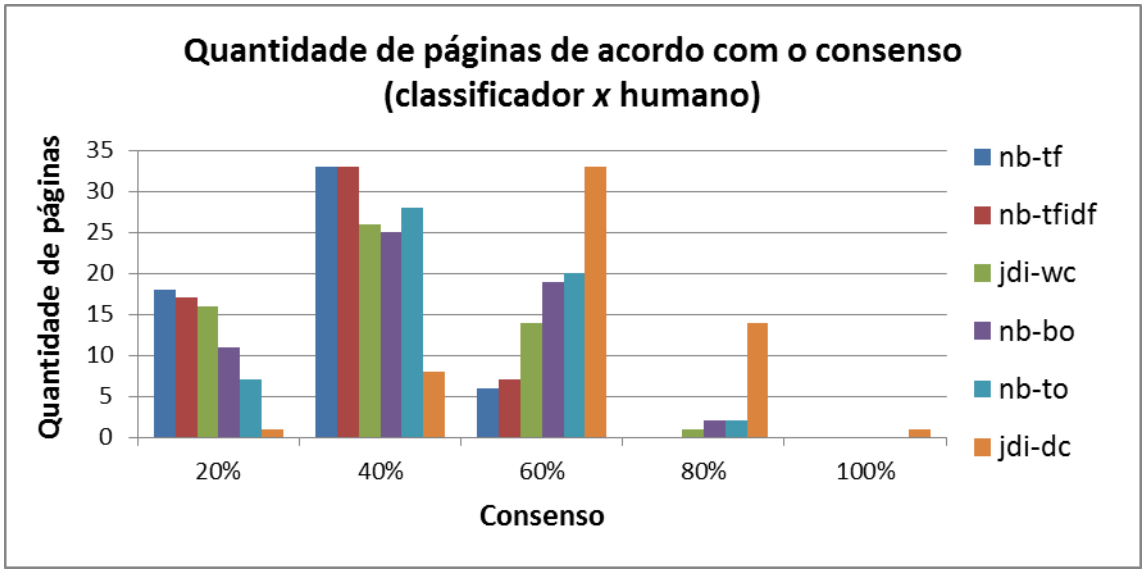

Figura 2. Quantidade de páginas em cada um dos valores de consenso entre a avaliação humana e cada um dos classificadores.

A última medida, o consenso entre os avaliadores, representa a porcentagem de avaliadores que assinalaram a mesma categoria para cada página. Na Tabela 2 é exibido o consenso médio entre os avaliadores para as cinco categorias mais escolhidas. Observa-se que, em média, mais de 85% dos avaliadores concordam em atribuir a

mesma categoria às páginas web. Já para a quinta categoria de maior consenso, cerca de 22% dos avaliadores atribuíram a mesma categoria. A diferença, de quase 63 pontos percentuais, mostra uma grande queda da primeira para a quinta posição, entretanto ainda representa a opinião difusa dos avaliadores, já que o consenso atingido é bem maior do que a escolha de uma categoria ao acaso (1/19 = 5,26%) [Santini 2008]. Os resultados são similares aos apontados por Santini (2008), que mostrou que um único rótulo não é o suficiente para corresponder às expectativas dos usuários.

Tabela 2. Média do consenso entre os avaliadores

	Média	Mín. - Máx.
1ª posição	**85,62%**	**50% - 100%**
2ª posição	**59,90%**	**14,29% - 100%**
3ª posição	**41,29%**	**14,29% - 71,43%**
4ª posição	**29,45%**	**0% - 71,43%**
5ª posição	**22,78%**	**0% - 57,14%**

Os três resultados preliminares (revocação, consenso entre avaliadores e classificadores, e consenso entre avaliadores) indicam a necessidade de se considerar multirrótulos para páginas web com conteúdo de saúde para corresponder às expectativas dos consumidores de informação sobre saúde.

Apesar dos resultados interessantes e promissores da avaliação da percepção humana de multirrotulação de páginas web de saúde em português brasileiro, o experimento realizado é preliminar e conta com algumas limitações.

A primeira delas é em relação à possibilidade de escolha de mais de uma categoria para cada página web pelos avaliadores. Com mais categorias por páginas, maior é a probabilidade de mais avaliadores terem escolhido a mesma categoria, contribuindo para que o consenso a partir da segunda categoria mais votada não sofresse uma queda grande em relação à primeira.

A quantidade de páginas selecionada, bem como a quantidade de avaliadores e o perfil dos mesmos também são fatores limitantes dos resultados. Com uma amostra pequena de páginas web é possível que haja perda de informações acerca de características diferentes de páginas web pertencentes às mesmas categorias, diminuindo a confiabilidade para extrapolar os resultados obtidos para um conjunto de páginas web maior. Devido à limitação na quantidade de páginas por categoria na base de dados coletada e a algumas páginas já não existirem, houve um limite de três páginas web por categoria, para haver uma uniformidade na quantidade de páginas por categoria, totalizando, assim, 57 páginas web para o experimento.

A amostra de avaliadores também pode não ser significativa para concluir que o comportamento encontrado seja para qualquer tipo de usuário. Apenas 21 pessoas participaram do experimento, e todos têm pelo menos nível superior completo, o que não representa com fidelidade uma amostra real de um público leigo que utiliza um buscador para procurar por informação sobre saúde na web.

4. Conclusões

A sugestão automática de multirrótulos para as páginas web de saúde colabora com recuperação de informação por pessoas leigas. De fato, houve consenso entre os avaliadores para as cinco categorias mais relevantes de uma página, mapeando satisfatoriamente a percepção dos avaliadores. Com isso, a utilização de uma ferramenta que sugira categorias automaticamente para páginas web de saúde, a partir da análise de seu conteúdo textual, pode direcionar melhor os consumidores de informação em saúde em suas buscas na web, satisfazendo melhor suas expectativas. Esta ferramenta pode ser acoplada diretamente a um buscador de conteúdos na web, que diferencie primeiramente os conteúdos de saúde e então indique categorias para as páginas web desta área.

Entretanto, novos estudos com quantidade maior de páginas e diferentes perfis de avaliadores são necessários para ser possível localizar fenômenos globais na web. Estudos no sentido de mapear o vocabulário médico em termos populares também podem melhorar o processo de multirrotulação de páginas web de saúde.

Agradecimentos

Os autores agradecem à CAPES-DS, à Grant NIH/ Fogarty 5D4 3TW007015-07 (PI : Dra Lucila Ohno Machado, Diretora Brasileira: Profa. Dra. Heimar de Fátima Marin), ao CNPq (Processo 559931/2010-7), à FAPESP (Processo 2010/15691-0) e ao Instituto Nacional de Ciência e Tecnologia – Medicina Assistida por Computação Científica (INCT-MACC), pelo apoio financeiro.

Referências

Barbosa, A. F. [Ed.] (2011). Pesquisa Sobre o Uso das Tecnologias da Informação e da Comunicação no Brasil 2010. . Centro de Estudos sobre as Tecnologias da Informação e Comunicação (CETIC). http://www.cetic.br/publicacoes/, [accessed on Feb 14].

Breitman, K., Casanova, M. A. and Truszkowski, W. (2006). *Semantic Web: Concepts, Technologies and Applications*. 1. ed. Springer.

Fogg, B. J., Soohoo, C., Danielson, D. R., et al. (2003). How do users evaluate the credibility of Web sites? A study with over 2,500 participants. In *Proceedings of the 2003 conference on Designing for user experiences*. . ACM.

Fox, S. (2011). The Social Life of Health Information, 2011. . Pew Research Center's Internet & American Life Project. http://pewinternet.org/Reports/2011/Social-Life-of-Health-Info.aspx.

Humphrey, S. M., Névéol, A., Browne, A., et al. (2009). Comparing a rule-based versus statistical system for automatic categorization of MEDLINE documents according to biomedical specialty. *Journal of the American Society for Information Science and Technology*, v. 60, n. 12, p. 2530–2539.

John, G. and Langley, P. (1995). Estimating Continuous Distributions in Bayesian Classifiers. *In Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, p. 338–345.

Keselman, A., Browne, A. C. and Kaufman, D. R. (2008). Consumer Health Information Seeking as Hypothesis Testing. *J Am Med Inform Assoc*, v. 15, n. 4, p. 484–495.

Kunder, M. De (2012). The size of the World Wide Web. http://www.worldwidewebsize.com/index.php?lang=EN, [accessed on Feb 14].

Mancini, F., Falcão, Alex Esteves Jaccoud, Hummel, A. D., et al. (2010). Brazilian health-related content web search portal development. In *Proceedings of the 13th World Congress on Medical and Health Informatics (MEDINFO 2010)*. . IOS Press.

Mancini, F., Sousa, F. S., Teixeira, Fábio Oliveira, et al. (2010). Use of Medical Subject Headings (MeSH) in Portuguese for categorizing web-based healthcare content. *Journal of Biomedical Informatics*, v. 44, n. 2, p. 299–309.

Qi, X. G. and Davison, B. D. (2009). Web Page Classification: Features and Algorithms. *Acm Computing Surveys*, v. 41, n. 2.

Rosso, M. (2005). Using genre to improve web search. University of North Carolina.

Salton, G. and Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, v. 24, p. 513–523.

Santini, M. (2008). Zero, single, or multi? Genre of web pages through the users' perspective. *Information Processing & Management*, v. 44, n. 2, p. 702–737.

Schembri, G. and Schober, P. (2009). The Internet as a diagnostic aid: the patients' perspective. *Int J STD AIDS*, v. 20, n. 4, p. 231–233.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys (CSUR)*, v. 34, p. 1–47.

Sousa, F. S. (2011). Análise Comparativa de Métodos de Recuperação de Informação para Categorização de Conteúdos Web Relacionados à Saúde. Universidade Federal de São Paulo (UNIFESP).

Sousa, F. S., Mancini, F., Teixeira, F. de O., et al. (2012). Categorização automática de conteúdos web de saúde em português brasileiro com classificador bayesiano. *Journal of Health Informatics*,

Stvilia, B., Mon, L. and Yi, Y. (2009). A model for online consumer health information quality. *Journal of the American Society for Information Science and Technology*, v. 60, n. 9, p. 1781–1791.

Tang, H. and Ng, J. H. K. (10 nov 2006). Googling for a diagnosis--use of Google as a diagnostic aid: internet based study. *BMJ*, p. bmj.39003.640567.AE.

Toms, E. G. and Latter, C. (1 sep 2007). How consumers search for health information. *Health Informatics Journal*, v. 13, n. 3, p. 223–235.

Wolfram, D., Wang, P. and Zhang, J. (2009). Identifying Web search session patterns using cluster analysis: A comparison of three search environments. *J. Am. Soc. Inf. Sci. Technol.*, v. 60, n. 5, p. 896–910.

Zeng, Q., Kogan, S., Ash, N., Greenes, R. A. and Boxwala, A. A. (2002). Characteristics of consumer terminology for health information retrieval. *Methods of Information in Medicine*, v. 41, n. 4, p. 289–298.