Brazilian Journal of Information Science

AVALIAÇÃO MULTILÍNGUE E DE LOCALIZAÇÃO DE MOTORES DE BUSCA DE *SITES* E BUSCA POR PALAVRAS-CHAVE

Anas AlSobh
Ahmed Al Oroud
Mohammed N. Al-Kabi
Izzat AlSmadi

Yarmouk University
Jordan

RESUMO

Os motores de busca estão competindo para serem vistos como universais, coerentes e independentes do idioma. Em princípio, os usuários que buscam informação através da Internet devem obter informações consistentes independentemente da linguagem e das palavras que estão usando, e independentemente da língua correspondente aos documentos pertinentes. No entanto, a linguagem deve afetar a sequência ou a ordem de classificação dos resultados obtidos. Neste projeto, várias ferramentas são construídas para avaliar palavras e demonstrações em vários idiomas. Os resultados são avaliados e comparados para possível correlação. Outra ferramenta é construída para rastrear *sites* de diferentes idiomas e locais, a fim de avaliar vários aspectos desses *sites*. Os resultados de ambos os estudos mostraram que, embora pareça que os motores de busca populares fazem progressos muito bons para a construção de motores independentes de idioma e local, no entanto, existem algumas limitações e situações em que a busca por resultados pode ser tendenciosa em direção à popularidade da linguagem e/ou localização do *site*.

Palavras-Chave: Recuperação da Informação; Motores de Busca; Processamento de Linguagem Natural; Tradução; Texto Correspondente; Pesquisa de Idiomas.

INTRODUÇÃO

A sobrecarga de informação é uma preocupação contínua para pesquisadores. Usuários em muitos casos são oprimidos pela quantidade de informação recuperada como resultados de suas pesquisas em motores/robôs de busca. Usuários da Internet em todo o mundo navegam buscando por informação relativa aos seus interesses. As principais ferramentas utilizadas para buscar por tais

BJIS, Marília (SP), v.4, n.1, p.2-19, jan./jun. 2010. Available in: < http://www2.marilia.unesp.br/revistas/index.php/bjis/index>. ISSN: 1981-1640

informações são os motores de busca. Motores de busca mantêm os bancos de dados atualizados por rastreamento contínuo através da Internet, para coletar e indexar todas as páginas da rede, documentos e conteúdos de *sites*. Para aumentar sua popularidade, grandes motores de busca estão evoluindo continuamente para abranger serviços em linguagem multilíngue. Por exemplo, no caso da língua árabe, o *Google* disponibilizou novos serviços que incluem: *Google Translate*, *Suggest* e *Google Ejabat* para responder a perguntas em árabe, preenchimento automático, o *Google Zeitgeist*, *Translated Search*, *Tashkeel* etc.

O Google como motor de busca principal para muitos usuários no mundo está evoluindo continuamente, melhorando e expandindo as ferramentas do *site* para abranger diferentes utilidades e para atingir usuários no mundo todo usando linguagens distintas. Usuários de outros idiomas, que não o inglês, realizam suas buscas usando palavras-chave em inglês ou usando palavras-chave de sua própria língua. Os resultados obtidos podem não corresponder exatamente com o que se queria inicialmente. Isso pode ser justificado afirmando que o usuário que busca usando uma palavra-chave em um determinado idioma, está interessado em obter os primeiros resultados naquela língua específica. Também usuários que buscam de um local específico podem querer obter os primeiros resultados com páginas relevantes de seu próprio país ou área do que as de outras línguas ou continentes. Contudo, em ambos os casos, eventualmente os resultados devem ser os mesmos, ou quase os mesmos. Indexadores de motores de busca devem isolar a camada da localização e linguagem do conteúdo real e documentos recuperados e indexados em suas próprias bibliotecas.

O objetivo desta pesquisa é propor a construção de indexadores que sejam independentes de linguagem. Vamos avaliar o *Google Translate*, juntamente com várias outras fontes de dicionários de código aberto, tais como o *Wordnet* (http://www.Wordnet.princeton.edu) etc. para comparar resultados de busca recuperados entre as palavras em árabe e os termos respectivos em inglês.

2 TRABALHOS RELACIONADOS

BJIS, Marília (SP), v.4, n.1, p.2-19, jan./jun. 2010. Available in: < http://www2.marilia.unesp.br/revistas/index.php/bjis/index>. ISSN: 1981-1640

Salton (1969) refere-se à *Cross-Language Information Retrieval* (CLIR) no final dos Anos 60 do Século 20, em que um dicionário multilíngue é usado para os documentos e pesquisas. Salton afirma que CLIR poderia ser tão eficaz quanto a recuperação de informação monolíngue.

Estudos de Al-Onaizan et al. (2002) apresentam uma solução para o problema de tradução denominado frases de entidade. Este é um problema difícil, já que tais frases são, em muitos casos, frases especiais relacionadas e não gerais. Como resultado, o usuário poderia não encontrá-las em dicionários gerais. Eles enfrentaram o problema denominado frases de entidade apresentando um novo algoritmo dedicado a traduzir árabe para inglês. Esse novo algoritmo adota diferentes abordagens para distintos tipos de frases de entidade, cujo processo de tradução é baseado em duas etapas principais. Na primeira etapa, uma lista ordenada de candidatos a tradução é produzida. Na segunda etapa, as traduções candidatas são recodificadas dependendo das pistas monolíngues. Posteriormente, traduções candidatas da primeira lista são reclassificadas, seus algoritmos transliteram e traduzem palavras árabes para o inglês e, em seguida, determinam se devem usar termos transliterados ou traduzidos em inglês.

Há tentativas para melhorar as pesquisas digitadas nas caixas dos motores de busca, e uma dessas tentativas por Loia et al. (2007) foi baseada em acrescentar semanticamente pesquisas semelhantes à pesquisa original, submetendo a pesquisa original para além das pesquisas semanticamente equivalentes, de forma que o motor de busca produzisse resultados diferentes. Os resultados são então unificados em uma lista filtrada. Essa abordagem objetiva ajudar os usuários a formular suas questões durante a sessão de busca, além de alcançar melhores resultados.

Diferentes abordagens de pesquisas semelhantes têm sido avaliadas por Balfe et al. (2005). Essas abordagens são classificadas em três categorias: termo baseado na semelhança de resultados baseado em métricas e comportamento de usuários em seleção de páginas relevantes.

BJIS, Marília (SP), v.4, n.1, p.2-19, jan./jun. 2010. Available in: < http://www2.marilia.unesp.br/revistas/index.php/bjis/index>. ISSN: 1981-1640

A representação da matriz é usada para encontrar os termos comuns entre diferentes pesquisas. O índice de relevância foi usado como uma medida baseada em abordagem de resultados. Os critérios de seleção baseados em abordagens usaram as medidas de seleção do usuário para páginas relevantes, visando encontrar a similaridade entre as pesquisas. Os resultados indicaram que o termo baseado em abordagens alcança os melhores resultados em termos de precisão e revocação.

Um gráfico para visualizar as pesquisas relacionadas que utilizam medida de similaridade de pesquisa híbrida para gerar grupos de pesquisas para cada consulta submetida foi proposta por Lin et al. (2004). O gráfico é gerado pela aplicação de agrupamento algoritmos de pesquisas e algoritmo TF-IDF para a construção do repositório de pesquisa, sobre as quais os *clusters* de pesquisas são construídos. Utilizou-se um questionário para medir a satisfação dos usuários em relação às novas abordagens para a sugestão de pesquisas relacionadas. Os resultados indicaram que, em termos de tempo, os usuários gastaram menos tempo formulando suas pesquisas usando o método gráfico. Guo e Bian (2008) propuseram um sistema de recuperação de informação multilíngue para documentos de patentes em inglês e japonês. Diferentes tradutores *web*, tais como *Google* e *Excite* são usados para traduzir as pesquisas. A tecnologia de indexação independente de linguagem é usada para processar as coleções de textos em muitas linguagens asiáticas. Os resultados indicaram que o método proposto atingiu resultados eficazes. Contudo, o sistema de recuperação de informação proposto não foi um *web-based*. Além disso, nenhum procedimento de *feedback* relevante foi usado.

Lianhau et al. (2009) construiu um sistema de recuperação de informação multilíngue denominado MARS. A criação do MARS é baseada na manipulação de uma coleção de documentos em *clusters* de conjuntos comparáveis, encontrando associações subjacentes entre as bases. O agrupamento de documentos foi realizado *off-line*; o agrupamento foi de fato a base para a recuperação de um documento comparável, multilíngue e relacionado de acordo com as pesquisas realizadas pelo usuário. MARS apoiou somente pesquisas simples em *GUI* e, portanto, foi menos apropriado para as pesquisas complexas.

BJIS, Marília (SP), v.4, n.1, p.2-19, jan./jun. 2010. Available in: < http://www2.marilia.unesp.br/revistas/index.php/bjis/index>. ISSN: 1981-1640

A eficácia de um sistema de recuperação de informação multilíngue capaz de lidar com quatro idiomas: inglês, chinês, japonês e koreano foi avaliada por Savoy (2005). A abordagem de combinação aproximada de tradução foi usada, em que os resultados indicam que a estratégia de tradução aproximada parecia aumentar a eficácia da recuperação para o chinês e japonês, mas não para o koreano. Esse estudo também abordou a estratégia de fusão de conjuntos de resultados gerados em diferentes línguas, cujo procedimento de fusão *Z-score* alcançou aproximadamente 5% a mais do que o método tradicional *round-robin*.

Uma ontologia orientada para a recuperação de informação de pesquisa multilíngue foi descrita por Nilsson et al. (2006). Um domínio de expansão de pesquisa específica e tradução foram usados. O processo de construção de ontologias do sistema foi composto por meio da coleta de conceitos específicos para a universidade, com o propósito de expansão da pesquisa. Sinônimos e hipônimos foram usados. Os termos correspondentes na língua-alvo foram usados para a pesquisa multilíngue. O sistema foi avaliado pelos usuários, contudo, o sistema proposto tem algumas deficiências no módulo de tradução.

Jang et al. (2002) aplicou o uso da recuperação de informação para pesquisas multilíngue em koreano para inglês e chinês. Um dicionário baseado no método de tradução foi usado. Um dicionário bilíngue foi usado para a tradução da pesquisa. Uma técnica de resolução de ambiguidade foi usada para remover termos desnecessários, bem como palavras inúteis que não têm efeito sobre o desempenho de recuperação. Para as pesquisas inglês-koreano o desempenho do sistema foi bem sucedido. Porém, para as pesquisas koreano-chinês o desempenho do sistema foi baixo. Verificou-se, também, a partir dos resultados obtidos que a tradução bilíngue tem seus próprios problemas e, portanto, o desempenho foi baixo.

Uma abordagem estatística foi usada para a tradução de pesquisas por Christof et al. (2005). Um dicionário bilíngue, bem como um monolíngue foi usado nos experimentos. É proposto um algoritmo que combina as medidas de associação com máquina de aprendizagem iterativa para cálculo de probabilidade. As probabilidades de tradução encontradas são usadas como peso do termo de pesquisa e, também, foram integrados vetores de espaço no sistema de

BJIS, Marília (SP), v.4, n.1, p.2-19, jan./jun. 2010. Available in: < http://www2.marilia.unesp.br/revistas/index.php/bjis/index>. ISSN: 1981-1640

recuperação. Os resultados mostraram que envolver uma abordagem incremental para a tradução de pesquisa, pode resultar em um melhor desempenho para a recuperação de informação entre multilinguagens.

A Teoria de Gráfico e o Método Padrão são técnicas propostas usadas pelos pesquisadores para resolver ambiguidades de tradução de pesquisa nos sistemas CLIR. Zhou et al. (2008) propôs um reforço hibrido de gráfico-padrão para melhorar o desempenho da tradução de pesquisa multilíngue na recuperação da informação. O método proposto se inicia traduzindo termos candidatos de um dicionário bilíngue. Por essa razão, várias traduções podem existir para o mesmo termo. Um padrão correspondente é usado no segundo passo para termos desconhecidos e ambíguos. Assim, todas as traduções que foram geradas no primeiro passo são direcionadas para um gráfico de representação, em que os termos com co-ocorrências são usados para se obter a melhor tradução. Os resultados da avaliação revelam uma melhoria promissora em relação aos métodos tradicionais.

A relação entre a sintaxe de um conjunto de palavras relacionadas poderia ser facilmente encontrada em diversos motores de busca (ex. *Google*, *Yahoo*). No entanto, motores de busca podem não considerar as relações semânticas que podem existir entre os conceitos. Danuska et al. (2009) propôs um método para encontrar a semelhança entre um conjunto de termos relacionados semanticamente. Um algoritmo de recuperação de padrões léxicos foi usado para representar relações semânticas comuns entre os termos (ex. *Google*, *Acquire*, *YouTube*). Um algoritmo de padrões sequenciais também foi usado para agrupar um conjunto de padrões de maneira apropriada, e então um vetor de características foi construído para encontrar a semelhança relacional entre os padrões recuperados. O teste que foi realizado com o método proposto e revelou uma melhoria em termos de desempenho e tempo de processamento.

Nos últimos anos, os motores de busca mudaram da recuperação de muitos dados irrelevantes para a recuperação de informação útil que podem ser analisadas por especialistas. Dessa forma, estamos atualmente nos movendo em direção à mineração de páginas recuperadas por qualquer mecanismo de busca na *Web*. Este tema foi discutido por Erinjeri et al. (2009), que destaca que o motor de busca do

BJIS, Marília (SP), v.4, n.1, p.2-19, jan./jun. 2010. Available in: < http://www2.marilia.unesp.br/revistas/index.php/bjis/index>. ISSN: 1981-1640

Google foi usado para explorar os relatórios de radiologia usando fontes de tecnologia livres e de código aberto. Uma ferramenta chamada *Radsearch* foi desenvolvida como parte de sua pesquisa, e construída sobre a infraestrutura do *Google*. Esta ferramenta permite que o Google recupere algumas páginas da *Web* relacionadas aos relatórios/laudos radiológicos.

Chew e Abdelali (2008) estudaram os efeitos de parentesco de linguagem sobre o desempenho multilíngue nos sistemas de recuperação de informação. Essa abordagem é usada para medir os efeitos da utilização de línguas semíticas nos sistemas de recuperação de informação multilíngue que incluem o árabe. Os resultados do estudo indicaram que o desempenho da CLIR aumentou extensivamente.

3 OBJETIVOS E ABORDAGENS

3.1 Pesquisa por Palavra-Chave

A fim de avaliar alguns temas relacionados à linguagem, construímos um pequeno banco de dados de palavras-chave mais visitadas em árabe em diversos países. As palavras-chave árabes mais visitadas armazenadas no banco de dados foram coletadas a partir do *Google*, *Alexa* e outros *sites* que monitoram informação por meio de palavras-chave mais visitadas e por país. Esses *sites* continuam monitorando os comportamentos dos usuários da Internet, e as palavras-chave que procuram/usam/buscam, ou seja, em outras palavras analisam o volume global de pesquisa sobre determinadas palavras-chave.

No *Google* isto é realizado através de várias ferramentas. Primeiramente, o *Google* sugere ou o preenchimento automático que é um método para mostrar aos usuários que iniciaram a digitação das letras e palavras às que correspondem às letras e palavras mais visitadas. A segunda e terceira fontes de informação para a maioria das palavras mais visitadas são: *Google Trend* (www.google.com/trends) e *Google Zeitgeist* (http://www.google.com/intl/enpress/zeitgeist/index.html). A quarta ferramenta do *Google* é a ferramenta de busca de palavras-chave *Sktool*

BJIS, Marília (SP), v.4, n.1, p.2-19, jan./jun. 2010. Available in: < http://www2.marilia.unesp.br/revistas/index.php/bjis/index>. ISSN: 1981-1640

(http://www.google.com/sktool), o qual fornece ideias de palavras-chave. A quinta ferramenta do *Google* é o *Google Insights for Search* (http://www.google.com/insights/search). Usando o *Google Translate* e outros dicionários inglês-árabe, o conjunto de palavras populares é traduzido para o inglês.

Para ambas, as árabes e suas respectivas palavras em inglês, o número de palavras e documentos relacionados de pesquisa são retornados. O número total de termos coletados excede seis mil termos para cada língua. Em alguns casos, algumas dessas palavras-chave são repetidas. Contudo, como essas palavras vêm de países distintos, elas são mantidas porque é esperado que mostrem diferentes resultados em termos de números de documentos correspondentes ou tráfego. Um rastreador e um robô são construídos para coletar dados automaticamente.

As pesquisas relacionadas ou buscas relacionadas no *Google* (Figura 1) mostram as palavras-chave que são relacionadas às atuais palavras-chave buscadas. Elas geralmente apresentam até oito resultados (geralmente mostradas no rodapé e às vezes no topo da página) do *Google* e ficam fixadas em todas páginas de retorno de busca. Tais pesquisas relacionadas podem depender de diversos parâmetros, e isso pode incluir o histórico de pesquisas que o *Google* mantém para os indivíduos (aquele que procura por *x*, também pode procurar por *y*). Também pode depender do processamento da linguagem natural e/ou semântica. Tráfego também é outro fator. Palavras-chave que aparecem com as pesquisas relacionadas têm alternado o filtro e são promovidas para aquela posição como resultado de volume de pesquisa.

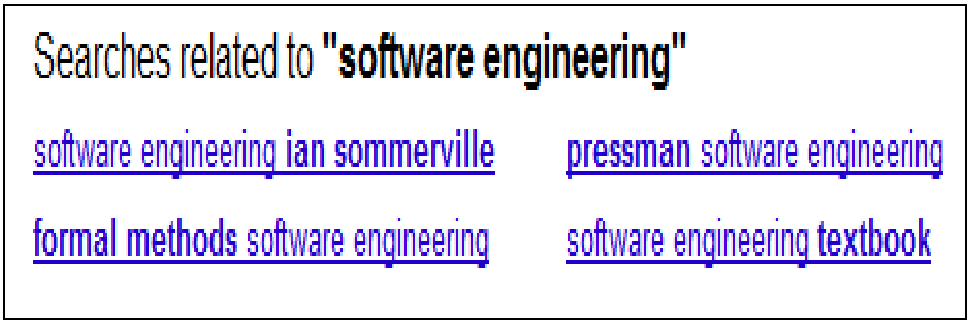

Figura 1: Google "Pesquisa Relacionada" para "Engenharia de Software" Palavra-Chave.

BJIS, Marília (SP), v.4, n.1, p.2-19, jan./jun. 2010. Available in: < http://www2.marilia.unesp.br/revistas/index.php/bjis/index>. ISSN: 1981-1640

3.2 Experimentos e Resultados

Um banco de dados é construído a partir de 6.589 palavras-chave, coletadas em árabe e inglês. Para cada palavra, em árabe e em inglês, o número de resultados retornados, ou seja, os resultados aproximados retornados ou documentos do mecanismo de busca se referem ao número de documentos que o motor de busca encontrou e ao número de pesquisas relacionadas ou "busca relacionada com" palavras-chave, ou seja, o número de termos de pesquisa relacionados às palavras-chave que o usuário usou, foram coletados. O objetivo foi estudar as variações e a dependência de busca de termos e documentos relacionados à linguagem usada para a busca. A Tabela 1 mostra a correlação de pesquisas relacionadas entre inglês e árabe. O *Google* retorna um número entre 0 e 8.

A Tabela 1 mostra que embora aquelas sejam palavras-chave populares em árabe, o *Google* encontrou mais palavras relacionadas com os termos relevantes em inglês. No entanto, na Tabela 2 o número de documentos relacionados ou recuperados é um pouco semelhante.

Para ver o outro lado da questão, 1.688 palavras-chave foram selecionadas através do *Google Sktool*, *Google Trend* e *Alexa*. Essas são as palavras mais usadas mundialmente. A Tabela 3 mostra o número de palavras de pesquisa relacionadas.

Tabela 1 - "Pesquisas relacionadas" comparação entre inglês e árabe para palavras-chave populares em árabe.

% de palavras-chave em inglês que retornam mais pesquisas relacionadas	% de número igual de pesquisas	% de palavras-chave em árabe que retornam mais pesquisas relacionadas
41.44	36.05	22.43

Tabela 2 - "Documentos recuperados" comparação entre inglês e árabe para palavras-chave populares em árabe.

% de palavras-chave em inglês que retornam mais documentos relacionados	% de número igual de documentos	% de palavras-chave em árabe que retornam mais documentos relacionados

BJIS, Marília (SP), v.4, n.1, p.2-19, jan./jun. 2010. Available in: < http://www2.marilia.unesp.br/revistas/index.php/bjis/index>. ISSN: 1981-1640

51.15	0.05	48.72

Tabela 3 - "Pesquisas relacionadas" comparação entre inglês e árabe para palavras-chave populares em inglês.

% de palavras-chave em inglês que retornam mais pesquisas relacionadas	% de número igual de pesquisas	% de palavras-chave em árabe que retornam mais pesquisas relacionadas
51	39.4	8.6

Tabela 4 - "Documentos recuperados" comparação entre inglês e árabe para palavras-chave populares em inglês.

% palavras-chave em inglês que retornam mais documentos relacionados	% de número igual de documentos	% palavras chave em árabe que retornam mais documentos relacionados
90.38	6.7	2.92

A Tabela 4 mostra o percentual de documentos recuperados entre o árabe e o inglês. Mostra também que mais de 90% dos documentos recuperados em inglês são maiores do que os recuperados em árabe. Somente menos de 3% das palavras em árabe recuperaram mais documentos.

A Figura 2 mostra a diferença entre a quantidade de documentos recuperados em inglês e sua tradução em árabe.

BJIS, Marília (SP), v.4, n.1, p.2-19, jan./jun. 2010. Available in: < http://www2.marilia.unesp.br/revistas/index.php/bjis/index>. ISSN: 1981-1640

English Word or Phrase	E-Retreived docs	Arabic Word or Phrase	A-Retreived docs
mortgage calculator	40,900,000	حاسبة التمويل العقاري	13,500
microsoft project	89,500,000	مشروع مايكروسوفت	765,000
ms project	285,000,000	مشروع السيدة	1,030,000
payroll	29,900,000	الرواتب	1,370,000
amortization calculator	3,080,000	آلة حاسبة اهتلاك	79
mortgage payment calculator	36,600,000	دفع الرهن العقاري حاسبة	1,430
interest calculator	45,200,000	آلة حاسبة الفائدة	23,200
content management system	867,000,000	نظام إدارة المحتوى	896,000
accounting software	35,400,000	برامج المحاسبة	1,480,000
home loan calculator	35,100,000	حاسبة قرض المنزل	49,400
project management software	94,100,000	برمجيات إدارة المشاريع	1,640,000
sales jobs	110,000,000	مبيعات وظائف	526,000
crm software	12,500,000	برمجيات إدارة علاقات العملاء	85,100
content management	202,000,000	إدارة المحتوى	4,950,000
amortization table	457,000	جدول السداد	101,000
peachtree	5,560,000	بشتر	47,700
loan payment calculator	16,300,000	تسديد القرض حاسبة	1,710
bulk email	23,500,000	البريد الإلكتروني السائبة	8,080
timesheet	1,770,000	الجدول الزمني	356,000
inventory management	28,800,000	إدارة المخزون	335,000

Figura 2: "Número de documentos recuperados" entre palavras-chave em inglês e árabe.

Espera-se que algoritmos dos motores de busca priorizem documentos recuperados baseados em diversos fatores, tais como tráfego ou popularidade. Isso pode explicar a razão pela qual as palavras em árabe pode não recuperar documentos na mesma ordem que aqueles em inglês, ou seja, as mesmas palavras traduzidas, pois isso reflete a popularidade das palavras em um determinado país ou região. No entanto, isso não deveria afetar, em grande medida, o número de documentos recuperados. A Tabela 4 indica que palavras populares no mundo têm um número muito pequeno de documentos recuperados em árabe.

Tabela 5 – "Pesquisas relacionadas" comparação entre inglês e outras línguas para palavras-chave populares em inglês.

% de palavras-chave em inglês que retornam mais pesquisas relacionadas	% de número igual de pesquisas	% de palavras-chave em alemão que retornam mais pesquisas relacionadas
32.69	50.96	15.38
% de palavras chave em inglês que retornam mais pesquisas relacionadas	% de número igual de pesquisas	% de palavras-chave em francês que retornam mais pesquisas relacionadas
67.3	30.7	0
% de palavras-chave em inglês	% de número igual de	% de palavras chave em

BJIS, Marília (SP), v.4, n.1, p.2-19, jan./jun. 2010. Available in: < http://www2.marilia.unesp.br/revistas/index.php/bjis/index>. ISSN: 1981-1640

que retornam mais pesquisas relacionadas	pesquisas	chinês que retornam mais pesquisas relacionadas
48	29.8	20.2

Tabela 6 – "Documentos recuperados" comparação entre inglês e outras línguas para palavras populares em inglês.

% de palavras chave em inglês que retornam mais documentos relacionados	% de número igual de documentos	% de palavras chave em alemão que retornam mais documentos relacionados
83.65	6.73	9.62
% de palavras chave em inglês que retornam mais documentos relacionados	% de número igual de documentos	% de palavras chave em francês que retornam mais documentos relacionados
74	12.5	13.5
% de palavras chave em inglês que retornam mais documentos relacionados	% de número igual de documentos	% de palavras chave em chinês que retornam mais documentos relacionados
89.4	0	10.6

Para resumir, a Tabela 7 mostra as buscas de pesquisas relacionadas com o número de documentos recuperados entre as cinco línguas. As porcentagens são apresentadas em relação ao inglês, ou seja, o foco da Tabela 7 é apenas sobre as percentagens em relação ao idioma inglês.

Tabela 7 – Pesquisa de busca relacionada ao número de documentos recuperados entre as diferentes línguas relativas ao idioma inglês.

Língua	Pesquisas de busca relacionadas	Número de documentos recuperados
Árabe	8.6	2.92
Alemão	15.38	9.62
Francês	0	13.5
Chinês	20.2	10.6

A Tabela 7 mostra que em ambas as buscas de "pesquisas relacionadas" e "número de documentos recuperados" indicam claramente que o inglês está dominando a Internet em relação aos outros quatro idiomas selecionados.

Existem dois papéis principais da linguagem nos *sites* e seus usuários. O impacto do número de (nativos) falantes de um idioma, determinam o número de *webhosts* nesse idioma, bem como o impacto do número de *webhosts* em certo idioma no número de *hyperlinks* conectando de/entre *sites* daquele idioma. O

BJIS, Marília (SP), v.4, n.1, p.2-19, jan./jun. 2010. Available in: < http://www2.marilia.unesp.br/revistas/index.php/bjis/index>. ISSN: 1981-1640

número de *sites* e leitores ou espectadores podem se beneficiar um com ou outro. O grande número de *sites* existentes na Internet, em uma língua específica pode contribuir para aumentar a popularidade daquele idioma. Por outro lado, um idioma, como o inglês, com um grande número de falantes propiciará melhor oportunidade e mais tráfego para os *sites* com esse idioma.

A conexão de e para um *site* é outro grande fator que afeta a popularidade de qualquer *site*. Isso também está diretamente relacionado à popularidade da língua/idioma e do número de falantes nativos. Em inglês em particular, a maioria dos falantes não são nativos e existem muitos *sites* no mundo todo que são escritos em duas línguas: a língua nativa e a língua inglesa.

3.3 Métricas da Popularidade

A fim de correlacionar a relação entre a língua e o país do *site* de um lado, com a sua popularidade de outro lado. Uma ferramenta foi desenvolvida para calcular os *inlinks* e os *outlinks* dos 10 *sites* mais populares de seis países selecionados baseados em sua língua. Esses seis países são: EUA para a língua inglesa, Alemanha para a língua alemã, Espanha para a língua espanhola, China para a língua chinesa, França para a língua francesa e Egito para a língua árabe.

Usando *Alexa.com* os 10 (dez) *sites* mais visitados desses seis países foram selecionados e seus *inlinks* e *outlinks* foram coletados. Nossa ferramenta desenvolvida para medir *inlinks* e *outlinks* utilizou diversos algoritmos para o pré-processamento, com o propósito de diminuir ou iluminar muitos *links* irrelevantes ou redundantes para *sites* que não afetam as métricas coletadas em uma grande extensão. Exemplo dessas páginas *web* ou componentes iluminados são aquelas páginas que são automaticamente geradas por ferramentas de *web design* e, consequentemente, serão vistas em todos os *sites*. A Tabela 8 apresenta os resultados obtidos destes *sites* selecionados. Zero *links* zero em alguns *sites* indica uma mudança de itinerário do *site*, tais como o www.msn.com que é convertido para www.bing.com.

BJIS, Marília (SP), v.4, n.1, p.2-19, jan./jun. 2010. Available in: < http://www2.marilia.unesp.br/revistas/index.php/bjis/index>. ISSN: 1981-1640

Tabela 8 - Métricas da popularidade para os 10 *sites* mais populares nos 6 países selecionados.

USA		France	
OutLink	**Inlink**	**OutLink**	**Inlink**
56	7320	78	1247
36	3266	**159**	**2847**
159	2847	**16**	**329**
745	7485	**745**	**7485**
1728	3249	**56**	**7320**
331	2355	**159**	**2847**
369	628	421	896
770	1217	**0**	**916**
16	329	6	274
371	1120	58	737
Egypt		**China**	
OutLink	**Inlink**	**OutLink**	**Inlink**
48	27	44	5678
36	**3266**	77	765
159	**2847**	56	6112
56	**7320**	123	682
745	**7485**	**56**	**7320**
149	480	712	2516
126	687	511	1621
16	**329**	233	867
369	**628**	82	657
37	**1604**	511	1543
Spain		**Germany**	
OutLink	**Inlink**	**OutLink**	**Inlink**
32	457	392	1370
159	**2847**	**56**	**7320**
16	**329**	**159**	**2847**
56	**7320**	**745**	**7485**
745	**7485**	722	1180
770	**1217**	**1728**	**3249**
159	**2847**	1254	2311
195	1261	**159**	**2847**
151	1052	1213	3126
0	**916**	432	678

Os resultados da Tabela 8 demonstram que, como esses, todos os *sites* populares estão alcançando grandes valores nos *inlinks* (também denominados *backlinks*). Contudo, o grande número em todos os países, tais como (7320 e 7845) são para o *Google* e o *YouTube*, que são *sites* populares na maioria dos países e línguas do planeta. Os números em negrito são para *sites* hospedados nos EUA com os valores coletados de outros países que não os EUA. Com exceção da China, todos os outros países estão recebendo cerca da metade dos seus *sites* populares

BJIS, Marília (SP), v.4, n.1, p.2-19, jan./jun. 2010. Available in: < http://www2.marilia.unesp.br/revistas/index.php/bjis/index>. ISSN: 1981-1640

dos EUA. Há uma correlação muito alta entre os resultados coletados a partir dos *backlinks* e a popularidade do *site*. No entanto, os *outlinks* não revelaram correlação positiva em todos os casos com a popularidade do *site*.

4 CONCLUSÃO E TRABALHO FUTURO

Essa pesquisa estudou os efeitos da localização do *site*, e da língua em sua popularidade. O trabalho também avaliou as diferenças entre os mesmos termos de pesquisa entre diferentes idiomas com base no país, e da própria língua. O inglês de fato ainda é a língua do mundo da Internet. Por outro lado, *sites* dos EUA, especialmente os populares, alcançam popularidade universal diferente de *sites* de outros países.

Os motores de busca oferecem muitos serviços para outras línguas, a fim de permitir igual oportunidade aos usuários da Internet, independentemente da língua ou localização. Entretanto, experimentos e estatísticas coletadas nesta pesquisa demonstraram que ainda existem muitas barreiras para realmente dar oportunidades iguais aos *sites*, independentemente de sua localização, país ou língua. Por outro lado, muitos *sites* internacionais têm uma versão em inglês. Em última instância, se espera que os motores de busca sejam concebidos de forma a tornar a linguagem, ou a localização, como características conectáveis que possam ser alternados em tempo de execução com a capacidade de traduzir todo o conteúdo do *site*, imagens, ícones etc. para o novo idioma de forma dinâmica.

No futuro, vamos propor uma nova estrutura para design de motores de busca que considerem eficientemente o idioma e o local do *site*. Um protótipo de motor de busca será construído e avaliado baseado no design proposto.

REFERÊNCIAS

ALMAS, Y.; AHMAD, K. LoLo: A system based on terminology for multilingual information extraction. In: CALIFF, M. E. et al. **Coling Association of Computational Linguistics 2006**. In: WORKSHOP ON INFORMATION EXTRACTION BEYOND THE DOCUMENT, Sydney, Australia, 2006. Sydney: ACL, 2006. p.56-65

AL-ONAIZAN, Y.; KNIGHT, K. **Translating named entities using monolingual and bilingual resources**. In: PROCEEDINGS OF THE 40TH ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS (ACL), 40., Philadelphia, 2002. Philadelphia: ACL, 2002. p.400-408

BALFE, E.; SMYTH, B. A comparative analysis of query similarity metrics for community-based web search. In: CASE-BASED REASONING RESEARCH AND DEVELOPMENT, 3620, 2005. **Proceedings…** p.63-77

CHEW, P.; ABDELALI, A. **The effects of language relatedness on multilingual information retrieval**: A case study with Indo-European and Semitic languages. In: PROCEEDINGS OF THE WORKSHOP ON CROSS-LANGUAGE INFORMATION ACCESS, 2008.

CHRISTOF, M.; BONNIE, J.; DORR, M. **Iterative translation disambiguation for cross-language information retrieval**. In: PROCEEDINGS OF THE 28TH ANNUAL INTERNATIONAL ACM SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL, 28., 2005. p.15-19

DANUSHKA, T. et al. **Measuring the similarity between implicit semantic relations from the web**. In: PROCEEDINGS OF THE 18TH INTERNATIONAL CONFERENCE ON WORLD WIDE WEB, 2009. Madrid, 2009.

DONG, Z. et al. A hybrid technique for English-Chinese cross language information retrieval. **ACM Transactions on Asian Language Information Processing (TALIP)**, v.7, n.2, p.1-35, 2008.

ERINJERI, J. P. et al. Development of a Google-based search engine for data mining radiology reports. **Journal Digit Imaging**, v.22, p.348-356, Apr. 2008.

GUO, W.; BIAN, S.; YUAN.T. Integrating query translation and text classification in a cross-language patent access system. In: PROCEEDINGS OF NTCIR-7 WORKSHOP MEETING, 2008. p.16-19

JACQUES, S. Comparative study of monolingual and multilingual search models for use with Asian languages. **ACM Transactions on Asian Language Information Processing (TALIP)**, v.4, n.2, p.163-189, 2005.

JANG, M. G. et al. **Simple query translation methods for Korean-English and Korean-Chinese CLIR in NTCIR experiments**. In: WORKING NOTES OF THE THIRD NTCIR WORKSHOP MEETING – PARTII: CROSS. 2002.

LIANHAU, L. et al. **Mars:** Multilingual access and retrieval system with enhanced query translation and document retrieval. In: THE 47TH ANNUAL MEETING OF ACL AND THE 4TH INTERNATIONAL JOINT CONFERENCE OF NLP (SW DEMO), 47., Singapore. Singapore: 2009. p.21-24

LIN, F. et al. **Query formulation with a search assistant**. In: ICADL, LNCS 3334. 2004. p.491-500

LOIA, V.; SENATORE, S. Customized query response for an improved web search. In: CASTILLO, O. (Ed.). **Theory advance and applied of fuzzy logic**. ASC 42, 2007. p.653-662

NILSSON, K.; HJELM, H.; OXHAMMAR, H. **Cross-language ontology-driven information retrieval in a restricted domain**. In: PROCEEDINGS OF THE 15TH NODALIDA CONFERENCE, 15., 2005. p.139-145

SALTON, G. **Automatic processing of foreign language documents**. In: PROCEEDINGS OF THE 1969 CONFERENCE ON COMPUTATIONAL LINGUISTICS. INTERNATIONAL CONFERENCE ON COMPUTATIONAL LINGUISTICS. ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. Morristown, 1969 p.1-28

ZHO, W.; YU, C.; MENG W. **A system for finding biological entities that satisfy certain conditions from texts**. In: CIKM'08. Napa Valley (CA), 2008. p.1281-1290

BJIS Brazilian Journal of Information Science

Anas AlSobh
Department of Computer Information Systems (CIS)
Faculty of Information Technology and Computer Sciences
Yarmouk University
Jordan

Ahmed Al Oroud
Department of Computer Information Systems (CIS)
Faculty of Information Technology and Computer Sciences
Yarmouk University
Jordan

Mohammed N. Al-Kabi
Department of Computer Information Systems (CIS)
Faculty of Information Technology and Computer Sciences
Yarmouk University
Jordan

Izzat AlSmadi
Department of Computer Information Systems (CIS)
Faculty of Information Technology and Computer Sciences
Yarmouk University
Jordan
E-mail: alsmadi@gmail.com