Uso de Expressões Temporais em Busca na Web: Uma análise através das sugestões de consulta

Augusto B. Corrêa[1], Edimar Manica[1 2], Renata Galante[1], Carina F. Dorneles[3]

[1]Instituto de Informática – Universidade Federal do Rio Grande do Sul (UFRGS)
Caixa Postal 15.064 – 91.501-970 – Porto Alegre – RS – Brasil
[2]Campus Avançado Ibirubá – Instituto Federal do Rio Grande do Sul (IFRS)
Rua Nelsi Ribas Fritsch, no 1111, Bairro Esperança – 98200-000 – Ibirubá – RS – Brasil
[3]Depto de Informática e Estatística – Universidade Federal de Santa Catarina (UFSC)
Caixa Postal 476 – CEP 88049-900 – Florianópolis – SC – Brasil
{abcorrea, emanica, galante}@inf.ufrgs.br, dorneles@inf.ufsc.br

Abstract. *This paper describes a study about Web Query Logs that identifies the way that users express temporal information in Web queries. The Web queries with temporal expressions have been analyzed on the following items: distribution in topics; mean size; most frequent queries; and, most frequent terms. To reach this goal, we have implemented a tool written in Python to get the query log from query suggestions.*

Resumo. *Este artigo descreve um estudo que analisa logs de consultas a fim de identificar a forma como os usuários expressam informações temporais em consultas na Web. As consultas com expressões temporais foram analisadas sobre os seguintes aspectos: distribuição em tópicos; tamanho médio, consultas mais frequentes e termos mais frequentes. Para atingir esse objetivo, foi implementada uma ferramenta em Python para coletar o log das consultas a partir de sugestões de consulta.*

1 Introdução

Os motores de busca mantém arquivos de *log* onde armazenam informações sobre a interação dos usuários. São armazenadas informações sobre as atividades de navegação (*clicks*) e de busca. Esses arquivos de *log* são úteis tanto para entender a estratégia de busca dos usuários, quanto para melhorar as sugestões de consulta [Wen e Zhang 2003] e a qualidade dos resultados retornados [Joachims 2002]. A análise de *logs* com foco na atividade de busca é conhecida como Análise de Logs de Busca (*Search Log Analysis*) [Trevisam et al. 2012].

Páginas web descrevem vários tópicos, tais como conferências, esportes, política e entretenimento. A maioria destes eventos mudam ao longo do tempo. A Escola Regional de Banco de Dados, por exemplo, ocorre todo ano. As olimpíadas ocorrem a cada quatro anos. A comunidade de banco de dados tem dedicado um esforço significativo para permitir a indexação e a consulta a dados temporais nas décadas passadas [Weikum 2011, Li et al. 2010]. Atualmente, o uso de expressões temporais tem emergido em consultas Web uma vez que documentos Web também possuem informações temporais [Manica et al. 2012]. Com isso, saber como o usuário expressa

sua necessidade de informação temporal é essencial para melhorar a qualidade dos resultados deste tipo de busca. Além disso, quando um usuário faz uma busca por uma pessoa famosa, os motores de busca exibem algumas informações dessa pessoa como foto, nome, nascimento, etc. Esse mecanismo não está disponível quando a busca é por uma expressão temporal. Quais informações deveriam ser exibidas neste tipo de busca? *Logs* de consultas com expressões temporais representam um recurso a ser analisado para responder essa pergunta.

A análise de *logs* tem sido um pouco limitada devido a falta de dados de usuários reais e a existência de questões éticas importantes [Bar-Ilan 2007]. Yoshinaga e Torisawa (2007) afirmam que é difícil para pessoas que não trabalham em empresas que possuem um motor de busca comercial obterem acesso a um grande conjunto de *logs* de consultas reais. Com isso, o presente trabalho pretende formar um *log* de consultas a partir das sugestões que um motor de busca comercial fornece quando submetida uma expressão temporal como consulta. Essa estratégia parte da observação que as sugestões fornecidas pelos motores de busca para uma consulta com uma expressão temporal são as consultas com aquela expressão temporal que mais ocorrem no *log* de consultas do próprio motor de busca. Por exemplo, ao submeter a expressão temporal "`17 de fevereiro`" como consulta ao motor de busca Bing[1], ele retorna as sugestões "`17 de fevereiro signo`", "`17 de fevereiro wikipedia`" e "`17 de fevereiro dia mundial do gato`". Isso significa que essas três sugestões são as consultas com a expressão temporal "`17 de fevereiro`" que mais ocorrem naquele motor de busca.

Este artigo descreve um estudo que analisa *logs* de consultas a fim de identificar a forma como os usuários expressam informações temporais em consultas na Web. O idioma adotado foi o português e o motor de busca utilizado foi o Bing. As principais análises realizadas sobre consultas com expressões temporais foram: distribuição em tópicos, tamanho médio, consultas mais frequentes e termos mais frequentes. Para atingir esse objetivo, foi implementada uma ferramenta em Python para coletar o *log* das consultas a partir de sugestões de consulta.

Este artigo está organizado como segue. A Seção 2 apresenta os principais conceitos temporais envolvidos no artigo. Na Seção 3 são apresentadas uma visão geral do trabalho desenvolvido e a configuração dos experimentos. Na Seção 4 é apresentada a análise dos dados coletados. A Seção 5 discute os principais trabalhos relacionados. Finalmente, na Seção 6, são apresentadas as considerações finais e as direções futuras.

1 http://br.bing.com/

2 Conceitos Básicos

Esta seção descreve os principais conceitos temporais necessários para a compreensão deste trabalho [Alonso et. Al 2007]:

- entidade temporal – é a descrição em um nível conceitual de um ponto no tempo, um evento ou um período de tempo;
- expressão temporal – é uma sequência de *tokens* que representa uma instância de uma entidade temporal;
- expressão temporal explícita – são expressões que descrevem diretamente uma entrada em uma linha de tempo, tal como uma data exata ou ano específico. Por exemplo, a expressão "`dezembro de 2013`" ou "`12 de janeiro de 2014`" em um fragmento de texto são expressões temporais explícitas e podem ser mapeadas diretamente para pontos em uma linha de tempo;
- expressão temporal implícita – são expressões que precisam de conhecimento predefinido (ontologias de tempo, por exemplo) para serem mapeadas para uma entrada em uma linha de tempo. Nomes de feriados e eventos específicos são típicos exemplos de expressões temporais implícitas. Por exemplo, a expressão "`Natal de 2013`" precisa ser mapeada para "`25 de dezembro de 2013`";
- expressão temporal relativa – são expressões temporais que representam entidades temporais que apenas podem ser mapeadas para uma entrada em uma linha de tempo em referência a uma expressão temporal explícita, implícita ou ainda ao momento em que o texto foi escrito. Por exemplo, a expressão "ontem" só pode ser mapeada com base no momento em que o texto foi escrito.

Manica, Dorneles e Galante (2012) classificam as consultas Web com informações temporais em dois tipos.

- seleção temporal (*temporal selection*) – consultas nas quais utiliza-se um predicado temporal para filtrar a consulta;
- saída temporal (*temporal output*) – consultas onde o usuário está interessado em saber qual o tempo em que um determinado evento ocorreu.

O presente trabalho restringe a análise a consultas Web de seleção temporal contendo expressões temporais explícitas, uma vez que este tipo produz maior número de consultas e logo, maior número de sugestões. Porém, como trabalho futuro se pretende trabalhar com os demais tipos de expressões temporais

3 Ferramenta e Configuração Experimental

Esta seção descreve como o *log* de consultas foi construído a partir das sugestões de consulta fornecidas por um motor de busca e as características do *log* de consultas construído. A Figura 1 apresenta a visão geral da ferramenta desenvolvida em Python utilizada para a construção do *log* de consultas. O usuário deve escolher entre um dos formatos de expressão temporal disponíveis pela ferramenta (por exemplo, "`DD de MM de AA`") e definir um valor inicial e um valor final. A ferramenta gera o conjunto de expressões temporais entre o valor inicial e o valor final de acordo com o formato definido. Cada expressão temporal gerada é submetida ao motor de busca Bing como uma consulta. O Bing, através da API (*Application Programming Interface*) `Qsonhs`[2], retorna as sugestões para a consulta submetida em um arquivo JSON (*JavaScript Object Notation*). Essas sugestões são extraídas através da biblioteca JSON Decoder[3], e enviadas para o *log* de consultas.

Figura 1: Visão Geral da Construção do log de consultas.

A Tabela 1 apresenta os formatos, valores iniciais e valores finais submetidos, sendo `DD` o dia em algarismos variando de 01 a 31, `MM` os meses escritos por extenso, `AA` o ano em algorismos variando de 1500 a 2100, `mm` o mês representado por algarismos variando de 01 a 12, `KK` representando a década variando entre 00 e 90 (apenas com números de final 0) e sendo `CC` o século em algarismos variando de 0 a 21. Neste trabalho foram submetidos apenas formatos de expressões temporais explícitas para consultas de seleção temporal, pois estas forneciam mais consultas e consequentemente um número maior de sugestões para a análise.

---

2 `Qsonhs`: é uma API do Bing para obter as sugestões para uma dada consulta. Basta executar a URL `http://api.bing.com/qsonhs.aspx?FORM=ASAPIH&q=CONSULTA`, substituindo o valor do parâmetro `q` por uma consulta. As sugestões são retornadas em um arquivo JSON.

3 `http://docs.python.org/2/library/json.html`

Tabela 1: Formatos, valores iniciais e valores finais submetidos.

Formato	Valor Inicial	Valor Final
DD de MM	01 de janeiro	31 de dezembro
DD de MM de AA	01 de janeiro de 1500	31 de dezembro de 2100
DD/mm	01/01	31/12
MM de AA	janeiro de 1500	dezembro de 2100
Século CC	Século 01	Século 21
Década de KK	Década de 10	Década de 90

A Figura 2 apresenta um exemplo de arquivo JSON retornado pela API `Qsonhs` para a consulta "`20 de setembro`". Como pode ser observado, foram retornadas 8 sugestões, sendo a primeira a própria consulta e a última "`20 de setembro feriado rs`". Não foi possível estabelecer um limite fixo de pesquisas diárias. Entretanto, notou-se um comportamento inusitado na API: a medida que eram realizadas mais pesquisas, o número de sugestões retornadas pela API eram menores que o número de sugestões fornecidas pelo Bing em consultas manuais. Para a correção de tal equívoco, utilizou-se um *sleep* – tempo em que o sistema ficou ocioso sem a utilização da API – de 120 segundos a cada 3600 pesquisas. Com este dispositivo, anulou-se a margem de distorção entre as pesquisas realizadas manualmente e as pesquisas realizadas através da API.

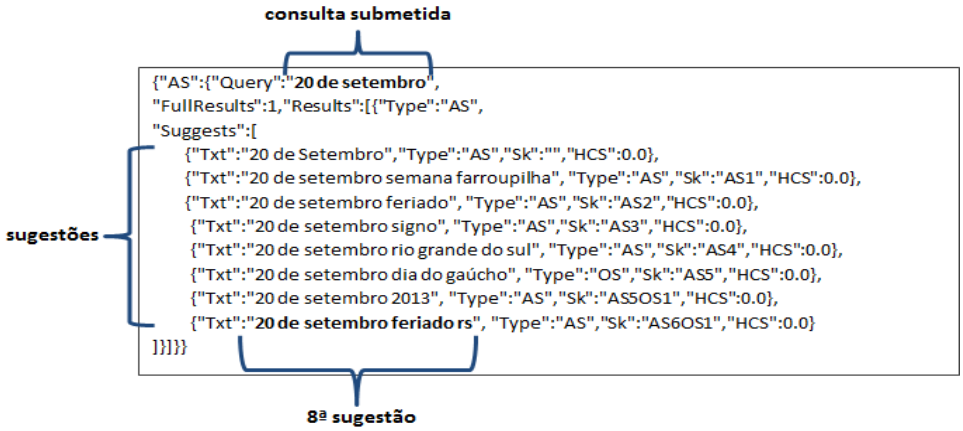

Figura 2: Exemplo de arquivo JSON retornado pela API Qsonhs

O *log* de consultas criado é um arquivo CSV (*comma-separated values*) contendo as seguintes colunas: (i) `Pesquisa` – expressão temporal que foi submetida na API Qsonhs; (ii) `Sugestão` – sugestão de consulta que foi retornada na API Qsonhs; (iii) `Data` – data e horário em que a pesquisa foi realizada; (iv) `Formato` - formato (Tabela 1) em que a consulta do termo Pesquisa foi realizada.

O *log* de consultas construído possui 529 Kbytes. A Tabela 2 apresenta o número de sugestões obtidas para cada formato.

Tabela 2: número de sugestões obtidas para cada formato.

Formato	Número de Sugestões
DD de MM	2425
DD de MM de AA	1127
DD/mm	2516
MM de AA	68
Século CC	42
Década de KK	62

4 Análise de dados

Esta seção descreve as quatro análises realizadas sobre o *log* de consultas com expressões temporais que foi construído conforme detalhado na seção anterior: *(i)* distribuição das consultas em tópicos; *(ii)* tamanho das consultas; *(iii)* termos mais frequentes; *(iv)* consultas mais frequentes. Essas análises foram realizadas sobre as sugestões de consulta armazenadas no *log*, excluindo os termos da consulta submetida. Por exemplo, para a sugestão "`15 de novembro proclamação da república`" obtida a partir da consulta "`15 de novembro`" apenas o fragmento "`proclamação da república`" é considerado nas análises. Esse fragmento é referenciado nessa seção como consulta.

4.1 Distribuição em Tópicos

O objetivo desta análise é verificar os tópicos que os usuários tem mais necessidade de informação temporal. As consultas foram classificadas em: entretenimento, datas comemorativas, notícias e sociedade, organizações, lugares, pesquisa, esportes e outros.

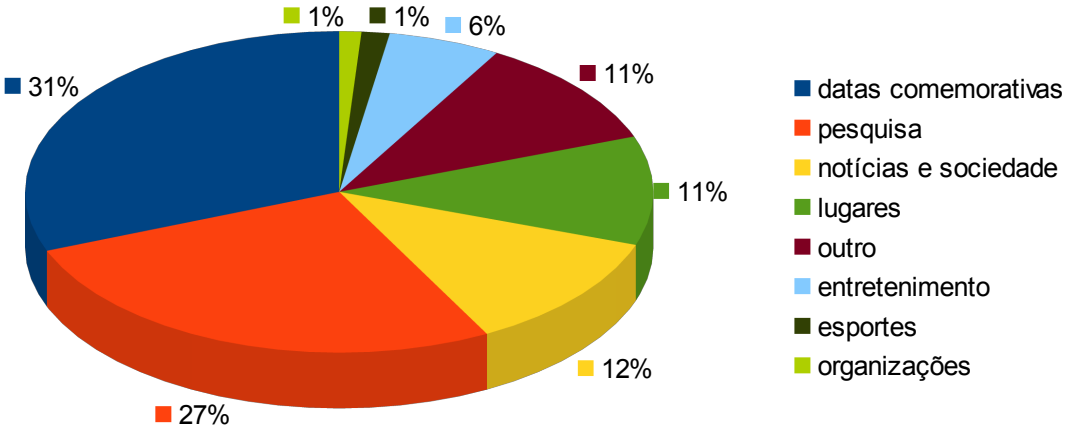

Figura 3: Classificação das consultas em tópicos

A Figura 3 ilustra os resultados obtidos (em ordem decrescente de frequência): datas comemorativas (31%), pesquisa (27%), notícias e sociedade (12%), lugares (11%), outro (11%), entretenimento (6%), esportes (1%) e organizações (1%). Dentro da categoria *outro* foram destacadas algumas subcategorias como, por exemplo: erro de digitação, finanças, compras, URL, computação e veículos, mas que não obtiveram frequência considerável para ser relevante no cenário da pesquisa.

4.2 Tamanho das consultas

O objetivo desta análise é identificar o tamanho das consultas com expressões temporais. Conforme ilustra a Figura 4, é possível observar que quase metade (47,3%) das consultas são compostas por dois termos. Em seguida, com um percentual bastante próximo, seguem as consultas com três termos (22,8%) e com um termo (18,4%). Por fim, houve um percentual bem pequeno de consultas com cinco termos (6,1%) e com quatro termos (4,3%). Nota-se também a presença de consultas de tamanho 6 (maiores consultas encontradas), embora não tenham atingido 1,0%.

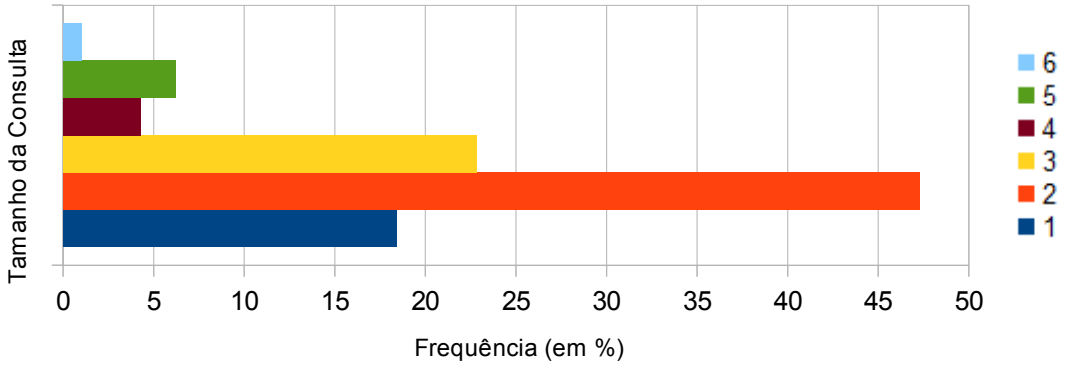

Figura 4: Tamanho da Consulta

4.3 Termos mais frequentes

O objetivo desta análise é identificar os termos que mais ocorrem em consultas com expressões temporais. Como pode ser observado na Tabela 3, os termos mais frequentes no *log* de consultas são: *dia, feriado, signo, nacional e brasil.* Nota-se também uma grande frequência de preposições (principalmente *de*, *do*, *e*, *da*), mas que não foram levadas em consideração uma vez que estas são *stopwords*, palavras frequentemente utilizadas mas com pouco significado isoladamente.

Tabela 3: Termos mais frequentes e suas de ocorrências no *log* de consultas.

Termo	Ocorrências	Termo	Ocorrências
Dia	629	Mundial	37
Feriado	409	Mundo	28
Signo	365	Carter	28
Nacional	45	Fim	25
Brasil	39	Vasco	25

4.4 Consultas mais frequentes

O objetivo desta análise é identificar as consultas que mais ocorrem com expressões temporais. As cinco consultas mais frequentes foram: *signo, feriado, é feriado, dia / dia de / dia do e qual signo.* Essas consultas representam as principais informações sobre uma expressão temporal que os usuários buscam na Web.

Tabela 4: Consultas mais frequentes e seu número de ocorrências

Consulta	Número de ocorrências
Signo	301
Feriado	172
É feriado	105
Dia / Dia de / Dia do	91
Qual signo	31

5 Trabalhos Relacionados

Nunes, Ribeiro e David (2006) analisaram o uso de expressões temporais em motores de busca e concluíram que as mesmas constituem uma fração muito pequena das pesquisas realizadas, porém dada a escalabilidade da Web representam um número considerável de usuários. Nunes, Ribeiro e David também se depararam com a análise de que a maioria das pesquisas que contém expressões temporais referenciam datas atuais ou passadas. Destoando de nossa pesquisa, os tópicos mais encontrados foram carros, esportes, notícias e sociedade, feriados, enquanto no presente trabalho foram datas comemorativas, pesquisa, notícias e sociedade e lugares, sendo que a categoria carros não atingiu 1%. Ressalta-se ainda que a análise realizada por Nunes, Ribeiro e David usou como base pesquisas de língua inglesa no motor de busca AOL.

Kato, Sakai e Tanaka (2013) analisaram as sugestões de consultas em mecanismos de busca Web, analisando três tipos de conjuntos de dados extraídos do

Bing: (1) *log* de consultas, (2) *log* de sugestões, (3) *log* de ações (*clicks*) dos usuários na página. De acordo com suas análises, foi possível elaborar os usos mais frequentes das sugestões de consulta pelo usuário: (1) quando a consulta original não é uma consulta frequente; (2) quando a consulta original contém só 1 termo; (3) quando não há ambiguidade na sugestão de consulta; (4) quando a sugestão é uma generalização ou uma correção da consulta original; (5) após o usuário clicar em diversos *links* na primeira página da consulta. Entretanto, não foram analisadas apenas pesquisas com expressões temporais, mas pesquisas de uma maneira geral. Ao contrário do presente trabalho, Kato, Sakai e Tanaka não efetuaram análises individuais sobre as sugestões de pesquisa – como classificar em tópico e tamanho – e apenas se limitaram a análise das sugestões de pesquisa como um todo.

Beltzel et al. (2007) realizaram uma análise temporal em um *log* de consultas da *American's Online* (AOL) dividido em tópicos – computação, música, carros, filmes, jogos, finanças pessoais, pornografia, saúde, entretenimento, viagens, websites americanos, casa e jardim, governo, esportes, compras e feriado – tentando investigar as mudanças nos padrões de consultas realizadas pelos usuários. Algumas categorias formaram certos padrões em sua frequência – seja ela anual, mensal, semanal ou diária. Por exemplo, as consultas classificadas na categoria *filmes* sofrem um acréscimo entre outubro e fevereiro, enquanto as pesquisas classificadas como *feriado* tendem a diminuir no meio da semana. Com o trabalho de Beltzel et al. torna-se possível analisar o comportamento do usuário de forma mais dinâmica e completa através das consultas dos usuários. Enquanto Beltzel et al. analisam o tempo em que a consulta foi submetida, o presente trabalho analisa a informação temporal que o usuário adicionou na consulta.

6 Conclusões

Com o aumento de dados temporais na Web, aumenta também a necessidade dos usuários de consultar tais dados. A fim de melhorar os resultados para esse tipo de consulta, é necessário saber como os usuários as expressam. Como os dados de *logs* de consultas não são disponibilizados pelos motores de busca, este trabalho criou um *log* de consultas através das sugestões de consultas fornecidas pelos motores de busca usando expressões temporais explícitas como consulta. Analisando este *log*, percebe-se que mais da metade das consultas são relacionadas a datas comemorativas e pesquisas na internet, além de que a maioria delas são curtas, tendo no máximo 3 termos em sua composição. Também, a análise de termos e consultas mais frequentes revela que ao

realizar uma busca por uma expressão temporal, o motor de busca poderia exibir o signo de quem nasce na data representada pela expressão temporal e se ela representa um feriado no contexto do usuário. Como trabalhos futuros se pretende expandir a análise para expressões temporais implícitas e relativas.

7 Referências

Alonso, O.; Gertz, M.; Baeza-Yates, R. A. (2007). "On the value of temporal information in information retrieval". In: SIGIR Forum, 41(2):35–41

Bar-Ilan, J. (2007). "Access to query logs - an academic researcher's point of view". In: Query Log Analysis: Social And Technological Challenges Workshop. 16th International World Wide Web Conference (WWW 2007)

Beitzel, S.; Jensen, E.; Chowdhury, A.; Frieder, O.; Grossman, D. (2007) . In "*Journal of the American Society for Information Science and Technology*", Volume 58 Issue 2, January 2007. Pages 166-178

Joachims, T.; (2002), "Optimizing Search Engines Using Clickthrough Data". In: Proceedings of the ACM Conference on Knowledge Discovery and Data Mining (KDD), ACM, 2002.

Kato, M. P.; Sakai, T.; Tanaka, K. (2013). In "*Information Retrieva*l", Volume 16 Issue 6, December 2013. Pages 725-746

Li, F.; Yi, K.; Le, W.; (2010). In: Top- queries on temporal data. VLDB J., 19(5):715–733.

Manica, E.;, Dorneles, C.; Galante, R.; (2012) "Handling temporal information in web search engines" SIGMOD Record (SIGMOD) 41(3):15-23 (2012)

Metzler, D.; Jones, R.; Peng, F.; Zhang, R. (2009). In "Improving search relevance for implicitly temporal queries". In *SIGIR '09 Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 700-701

Nunes, S.; Ribeiro, C.; David, G. (2008). In "Use of temporal Expressions in Web Search". ECIR 2008:580-584

Oliveira, J. P.; Edelweiss, N. (1994). "Modelagem de Aspectos Temporais de Sistemas de Informação". IX Escola de Computação, Recife.

Trevisan, M.; Barbu, E.; Barsanti, I.; Dini, L.; Lagos, N.; Segond, F.; Rhulmann, M.; Vald, E.; (2012) "Query log analysis with LangLog". In: EACL 2012:87-91

Weikum, G.; (2011). "Longitudinal analytics on web archive data: Its about time!". In: 5th Biennial Conference on Innovative Data Systems Research (CIDR2011), 2011. Wen, J; Zhang, H. J.; (2003). "Query Clustering in the Web Context".In: Clustering and Information Retrieval, Kluwer

Yoshinaga, N.; Torisawa, K. (2007). Open-Domain Attribute-Value Acquisition from Semi- Structured Texts. Workshop on Ontolex-07, [S.l.].