Quantifying Asymmetric Semantic Relations from Query Logs by Resource Allocation

Zhiyuan Liu, Yabin Zheng, and Maosong Sun

Department of Computer Science and Technology,
State Key Lab on Intelligent Technology and Systems,
National Lab for Information Science and Technology,
Tsinghua University, Beijing, China, 100084
{liuliudong,yabin.zheng}@gmail.com, sms@tsinghua.edu.cn
http://nlp.csai.tsinghua.edu.cn

Abstract. In this paper we present a bipartite-network-based resource allocation(BNRA) method to extract and quantify semantic relations from large scale query logs of search engine. Firstly, we construct a query-URL bipartite network from query logs of search engine. By BNRA, we extract asymmetric semantic relations between queries from the bipartite network. Asymmetric relation indicates that two related queries could be assigned different semantic relevance strength against each other, which is more conforming to reality. We verify the validity of the method with query logs from Chinese search engine Sogou. It demonstrates BNRA could effectively quantify semantic relations from We further construct query semantic networks, and introduce several measures to analyze the networks. BNRA is not only 'language oblivious' and 'content oblivious', but could also be easily implemented in a paralleled manner, which provides commercial search engines a feasible solution to handle large scale query logs.

Key words: semantic relations, query log, resource allocation, asymmetric

1 Introduction

With the development of Internet, search engine, such as Google, has become the most important tool to get information on the World Wide Web. Although it is not a perfect method to find what users want, most search engines calculate the relevance using keywords in documents and queries. As the only interface for users to access tremendous web pages, queries are one of the most important factors that affect the performance of search engines.

However, web pages returned from search engines are not always relevant to search intentions of users. An independent survey of 40,000 web users found that after a failed search, 76% of them will try to rephrase their queries on the same search engine instead of resorting to a different one [1]. Therefore, it is a non-trivial task for search engines to find better query representation of user search intentions in order to enhance search performance.

Search behavior of most users, including query submissions and URL selections, are meaningful. Therefore, queries convey implicit knowledge, concepts or meanings, which could be regarded as tags assigned to selected URLs by users [2]. Query logs, recording the history of click through behaviors by users from queries to selected URLs, may thereby contain tremendous user collaborative tagging information as a result of the wisdom of all users, and have attracted much work trying to extract useful information so as to improve search engine performance. Various tasks, such as query clustering [1, 3], classification [4, 5], recommendation [6], expansion [7] and reformulation [8], have been proposed to address the challenges from different perspectives. The common basis of these tasks is to quantify semantic relations of queries.

In a narrow sense, **semantic relations** are relations between concepts or meanings. Queries, regarded as the tags assigned by users to selected URLs, contain rich semantic relations, which imply a taxonomy of the language that people use to search for information [2]. Hence, it is essential to extract useful relations from query logs in order to improve search engine performance in various tasks mentioned above.

Most previous researches extracted semantic relations by defining a similarity function between queries based on substring matching of queries or intersection of selected URL sets. The main drawback of these methods is that the extracted relations are symmetric, which indicates the similarity function gives two queries the same relevance strength against each other. However in most instances two related queries should be assigned different relevance strength. For example, the relevance strength for query 'ipod'(a product of Apple Inc.) with 'apple' may be stronger than that for 'apple' with 'ipod': with query 'ipod', users likely want to get the information of the websites on the mp3 product ipod, therefore it is related with it manufacturer 'apple' strongly. While with query 'apple', users may have more complicated and extensive intentions, could be a fruit or an IT company, and thus not have equal strong relevance with 'ipod'. Hence, it is crucial to extract and quantify asymmetric semantic relations of queries.

In this paper we propose to apply a bipartite-network-based resource allocation(BNRA) method [9] to flexibly extract and quantify asymmetric semantic relations of queries for the first time. The method is originally applied to personal recommendations [10]. It is reported that, despite of simplicity, the method performs much better than most commonly used ranking method, such as global ranking method and collaborative filtering [9]. We also have got an initial but encouraging result using the method for query suggestions [11]. The work here follows the idea of [11]. BNRA method has three prominent features, namely asymmetric, parallelable and 'content oblivious'. In this paper, we verify the validity of the method for extracting semantic relations from query logs. We also analyze large query semantic networks constructed with the asymmetric relations.

2 Previous Work

There has been much related work on extracting query semantic relations, most of which is related to query clustering, classification, recommendation, expansion or reformulation, and is usually carried out on bipartite networks constructed from query logs with one node set containing only queries and the other URLs. Various methods compute query relations according to the similarity between returned documents [12, 13], selected documents [14, 6] or snippets of returned results [15]. Most of them reported satisfactory results, but unfortunately are not applicable for large-scale documents due to unacceptably massive calculations.

Beeferman and Berger [1] proposed a 'content oblivious' method to generate a list of related query formulations for input queries by merging most related queries or URLs in query logs alternately, where the relations are measured in terms of the number of overlapped neighbors in bipartite networks. Wen et al. [3] proposed a better-designed solution for query clustering by combining content-based and link-based clustering together and using four notions of query relations, i.e. keywords or phrases in queries, substring matching of queries, common selected URLs and the similarity of selected documents. A method based on association rule was also proposed to discover related queries from a set of search transactions or sessions, where each session includes a sequence of input queries by a single user in a certain time interval [16]. One apparent disadvantage of the method is that it could merely find related queries submitted by the same user while incapable to extract most related queries submitted by different users. Query relations can also be explored by mapping queries to predefined topic categories like Broder's [17] informational/navigational/transactional taxonomy [18], geographical-locality-based categories [19] or other artificial categories [4, 5]. Query classification brings great improvements to search engines, but on the other hand confines query relations into certain predefined categories. Baeza-Yates [20] described several relations between queries based on different information sources, i.e. keywords of queries, selected URL covers as well as hyperlinks or terms in selected web pages, and different semantic networks were defined based on these relations, among which the relations based on selected URL covers were qualified to be of the highest semantic strength.

A crucial common drawback of above methods is that the extracted query relations are symmetric, while asymmetric semantic relations are ubiquitous and more conforming to real world. The most relevant work to this notion was done by Baeza-Yates [2], where asymmetric query relations were extracted from Query-URL bipartite network based on selected URL covers. However, these asymmetric relations are associated from query q_i to q_j only when the URL set selected from q_i is completely covered by that from q_j, which restricts the extraction capability of asymmetric relations. BNRA, in contrast, is capable to extract and quantify asymmetric relations in a more natural and flexible manner. In subsequent sections, we will systematically investigate detailed properties of BNRA, including the recursive BNRA and its convergence, tunable parameters, etc.

3 BNRA Method

In order to implement BNRA, we need to construct a weighted Query-URL bipartite network from query logs. The click frequencies from queries to URLs suggest the matching degree between search intentions behind queries and semantics behind URLs. Hence, it is essential to assign weight to each edge between query and URL based on click frequency. Denoting the query set as $Q = \{q_1, q_2, ..., q_n\}$ and the URL set as $U = \{u_1, u_2, ..., u_m\}$, the bipartite network could be described by an $n \times m$ adjacent matrix $A = a_{ij}$, where $a_{ij} > 0$ if u_j is clicked under submitted q_i, indicating the click frequency, and $a_{ij} = 0$ otherwise.

3.1 Method Description

BNRA is elaborated as follows. To find related queries for query q_i based on the network and quantify their relevance strength, we initially assign resource value f_i to query q_i, indicating the semantic information kept by q_i. Afterwards, the resource-allocation is processed in two steps. Firstly, the resources in query nodes (Initially only q_i keeps resource.) are proportionally distributed, in terms of corresponding edge weight, to their neighbor URL nodes. Whereafter, the resources in URL nodes, are proportionally propagated to their neighbor query nodes in reverse. The final resources located in a subset of query nodes, denoting as R_i, are regarded as the distribution of semantic information of q_i and indicate the relevance strength between q_i and the queries in R_i. The relevance strength from query q_i to $q_j \in R_i$, denoting as r_{ij}, reads

$$r_{ij} = f_i \times s_{ij} \tag{1}$$

$$s_{ij} = \frac{1}{k(q_i)} \sum_{l=1}^{m} \frac{a_{il} a_{jl}}{k(u_l)} \tag{2}$$

where $k(q_i) = \sum_{j=1}^{n} a_{ij}$ and $k(u_l) = \sum_{j=1}^{m} a_{jl}$ are weighted degrees of query q_i and URL u_l. Denoting strength matrix as $S = (s_{ij})_{n \times n}$, and initial resource distribution in query set as the row vector $\mathbf{f}^{(0)} = (f_1, f_2, ..., f_n)$, the final resource distribution is $\mathbf{f}^{(1)} = \mathbf{f}^{(0)} \cdot S$. In matrix S, the ith row indicates the resource distribution in queries originated from query q_i after resource allocation. S has the property that the sum of all values in each row is equal to unity 1, namely, $\sum_{j=1}^{n} s_{ij} = 1, \forall i = 1, ..., n$.

3.2 Computational Complexity

BNRA introduces high efficiency in both space and time. Denoting the edge number as e and the maximum degree of queries or URLs as k_{max}, BNRA for all queries requires $O(nk_{max}^2)$ operations and simply $n \times m$ memory for storing a bipartite network. In contrast, the classical agglomerative clustering method [1] requires $O((n+m)k_{max}^2 + e(4k_{max}))$ operations and $n+m+n^2+m^2$ memory for storing query similarities and URL similarities besides a bipartite

network. An advanced method based on intersections of selected URL sets [2] requires $O(n^2 k_{max})$ operations. Therefore, BNRA is more efficient than methods mentioned. Moreover, compared to the agglomerative clustering method, BNRA could be implemented in a parallel manner with ease, which serves as a defining utility for commercial search engines in mining large-scale user logs. On the other hand, it could extract and quantify query relations more flexibly than the method proposed in [2].

3.3 Recursive BNRA and Its Convergence

It is a natural conjecture that the resource allocation process in BNRA could be executed recursively as $\mathbf{f}^{(t+1)} = \mathbf{f}^{(t)} \cdot S = \mathbf{f}^{(0)} \cdot S^t$, where $\mathbf{f}^{(t)}$ indicates the resource distribution after the tth iteration. Such approach might extend the method to a diffusion-like algorithm which would converge to a stable solution of equation $\mathbf{f}^* = \mathbf{f}^* \cdot S$ mentioned by Zhou et al. [21] without further discussion. Here, we elaborate more detailed analysis.

If we regard all queries as the states transiting from one to another according to the corresponding transition probabilities, and S as the transition probability matrix, the recursive resource allocation process is in nature a Markov process among queries [22]. According to Markov Process Theory, if the Markov chain is irreducible and aperiodic, there is a unique stationary distribution $\mathbf{f}^*$, and S^t converges to a rank-one matrix in which each row is the stationary distribution, that is

$$\lim_{t \to \infty} S^t = \mathbf{1} \cdot \mathbf{f}^* \tag{3}$$

where $\mathbf{1}$ is the column vector with each value equal to 1.

In practice, a query-URL bipartite network can be composed of one large component and many small components. The set of queries within one connected component is a communicating class thus the corresponding Markov chain is irreducible. The bipartite network component is connected based on complicated behaviors of users. Therefore, the Markov chain is aperiodic. As Eq. (3) suggests, originally from any query in one component, when reach the unique stationary state, the resource distribution $\mathbf{f}^*$ is uniform, only determined by the topology of the bipartite network and has nothing to do with the initial resource distribution. The feature indicates on one hand the recursive process can expand the related queries effectively; on the other hand it may lower the relevance strength with the original query and strengthen the effect of global popularity, which is a trade off between the relevance specificity and global popularity.

3.4 Tunable Parameters

Two tunable parameters may effect the performance of BNRA. One is the iteration number t. As mentioned above, the iteration times can effect the range of related queries and the resolving power on relevance strength between queries. The other parameter is the resource allocation strategy. A naive strategy is to

allocate the resource according to click frequency as shown in Eq. (1)(2). A more complicated form is

$$s_{ij} = \frac{1}{k(q_i)} \sum_{l=1}^{m} \frac{a_{il}^{\alpha} a_{jl}^{\alpha}}{k(u_l)} \tag{4}$$

$$k(q_i) = \sum_{j=1}^{m} a_{ij}^{\alpha}, k(u_l) = \sum_{j=1}^{n} a_{jl}^{\alpha} \tag{5}$$

where α is a tunable parameter controlling the force of click frequency on resource allocation, comprising the condition in Eq. (1)(2) when $\alpha = 1$.

Next, we will verify BNRA in query logs of Chinese search engine obtained from Sogou Labs and inspect the effect of the two parameters. Before coming to the details of the experiments and evaluation, we introduce the user log dataset in advance.

4 Experiment and Evaluation

4.1 Query Log Dataset

In our experiment, we use the query logs in the first week of March 2007 from Sogou Labs. Sogou Labs, founded by Chinese commercial search engine Sogou, consist of various web search resources of Sogou including query logs in one month, which can be accessed from `http://www.sogou.com/labs/`. There are 10,046,246 inquiry instances, 1,310,135 unique queries, 980,395 keywords and 4,055,171 unique URLs in this query log, where we refer *query* to a string submitted to search engines by a user which may contains one or more *keywords* delimited by white spaces, and *query instance* to one click behavior from a query to a URL. The number of keywords in queries mostly ranges from 1 to 3, and most of all keywords consist of 2 to 6 Chinese characters. Due to the shortness of keyword length, query relations extracted via keyword substring matching may be sparse and the performance will be greatly limited.

We apply BNRA and filter out the queries and URLs occurred only once in order to reduce noises. The constructed bipartite network contains 834,107 unique queries and 886,702 URLs. For each query, we assign resource $f_i = 100$, execute the resource allocation process with only one iteration and record top nine related queries. Table 1 shows some examples where related queries are listed according to the relevance strength in reverse order. For '小说网', '小说' (The English translations of these Chinese queries can be found in appendix, and hereinafter the same.) is positioned at the first place with strength 28.96. While for '小说', '小说网' is positioned at the last place with strength 1.44.

4.2 Evaluation

Most commercial search engines recommend queries they consider to be related to the original query as related search. We compare our method with recommended results by commercial search engines, i.e. Baidu (`www.baidu.com`) and

Table 1: Examples of related queries extracted by our method.

Query	Related Queries
小说网	**小说**, 玄幻小说, 小说阅读网, xiaoshuo, 言情小说, 小說, 免费小说, 中文小说网, 言情
小说	玄幻小说, 起点, 小说阅读网, 言情小说, 潇湘书院, 起点中文网, **小说网**, 幻剑书盟, 起点中文

Table 2: Recommended queries to '文学' from BNRA, Baidu and Google.

Source	Related Queries
BNRA	**榕树下**,玄幻小说,世纪文学,成人小说,小说,原创文学,文学小说,读书,天地文学
Baidu	世纪文学,文学屋,天翼文学,吾爱文学网,起点文学,文学家,成人文学,晋江文学网,晋江文学
Google	世纪文学,吾爱文学网,星辰变世纪文学,79文学网,天地文学,文学城,艳情文学,极品家丁世纪文学,文学殿堂,文学屋

Google (`www.google.cn`). As shown in Table 2, we compare the recommended queries of query '文学'. In most cases, the two search engines recommend those queries which contain the original query as substring. On the contrary, BNRA could extract related queries with no common substrings, which extends the scope widely. For example, the first query '榕树下' recommended by BNRA is the largest website of Chinese original literature, having no common substring with '文学'.

Users' perception indicates the performance of search engines to some extent. Therefore, we use editors' ratings to evaluate the performance of BNRA. We randomly select about 180 recommended queries and ask editors to rank these queries from 5 to 0, where 5 means very good and 0 means totally unrelated. All rating data can be accessed through `http://nlp.csai.tsinghua.edu.cn/~lzy/qr/rqj.zip`. In Fig. 1 we show the average scores of Baidu and BNRA with different iteration numbers. Despite of the disagreement among editors, the performance of BNRA is comparable with Baidu, which demonstrates the method is feasible and effective. Fig. 1 also suggests the loss of the specific relevance with original queries during the recursive resource allocation process. By Fig. 2, we show the average scores considering different numbers of recommended queries in the experiment with one iteration and the plot is skewed which is consistent with the decline of the relevance strength quantified by BNRA.

The agglomerative clustering method [1] was also performed on the dataset. Since the method requires huge memory, we compressed the bipartite network by filtering out the queries with unique click frequency lower than 10. The method iteratively merges the most related query pair and URL pair alternately until a termination condition applies. One reasonable termination condition proposed by Beeferman and Berger [1] is

$$\max_{q_i,q_j \in Q} \sigma(q_i, q_j) = 0 \text{ and } \max_{u_i,u_j \in U} \sigma(u_i, u_j) = 0$$

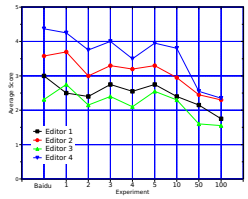

Fig. 1: Average scores of recommended queries by four editors.

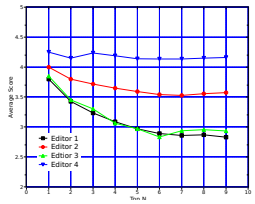

Fig. 2: Average scores of different numbers of recommended queries.

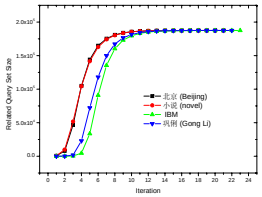

Fig. 3: Related query set size growth with iterations.

where $\sigma(*,*)$ indicates the fraction of common neighbors of two queries or URLs. However, it makes no sense for finding related queries because it leads to find the connected components where the queries distribute extremely imbalanced and most are in several large components. Besides, the agglomerative clustering method is time-consuming. In a PC with *Intel Duo* $2.80GHz$ CPU and 1.5GB memory, it spent about 400 minutes on the compressed bipartite network constructed by the queries with unique click frequency more than 10. For BNRA, however, it spent no more than 1 minute to deal with the bipartite network constructed by the queries occurred more than once. In addition, the optimal termination condition of hierarchical agglomerative clustering algorithms is not resolved efficiently [23], so it is hard for the agglomerative clustering method to find an optimal solution.

4.3 Parameter Effects

In this subsection, we inspect the effect on BNRA of iteration numbers and α in Eq. 4. In Fig. 3 we illustrate the size changes of related queries for several queries after each iteration until finding the whole connected components. We also show the changes of top five related queries of query '小说' during the first four iterations in Table 3 with no dramatic changes found for the top related queries.

Table 3: Top 5 related queries of query '小说' after 1 to 4 iterations. The values in the brackets are relevance strength.

Iteration	Related Queries
1	玄幻小说(23.8), 起点(6.5), 小说阅读网(5.2), 言情小说(4.9), 潇湘书院(1.9)
2	玄幻小说(28.6), 起点(6.7), 言情小说(4.8), 小说阅读网(3.3), 起点中文网(2.2)
3	玄幻小说(30.1), 起点(6.4), 言情小说(4.6), 小说阅读网(2.4), 起点中文网(2.2)
4	玄幻小说(30.4), 起点(5.9), 言情小说(4.4), 起点中文网(2.1), 小说阅读网(2.0)

In order to track the changes during the iterations, we use Euclidean distance to measure the variation between two adjacent resource distributions. Four

queries' variation dynamics along with iterations is shown in Fig. 4. Each of them terminates until the variation is less than 0.1. In in Fig. 5, we also illustrate the variation in each iteration of several semantic-free query pairs within one connected component of bipartite network, which indicates the trend towards convergence.

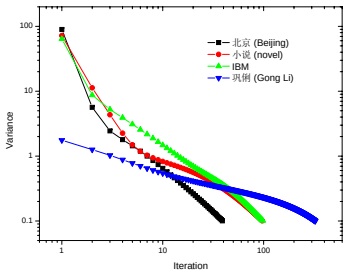

Fig. 4: Log-log plots of four queries' variation dynamics along with iterations.

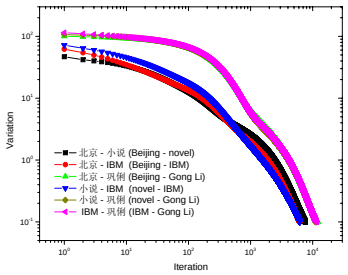

Fig. 5: Log-log plots of the variation along with iterations of several semantic-free query pairs.

Another tunable parameter is the α in Eq. 4 which effects the resource allocation strategy. In Fig. 6, we show the resource distributions of query '北京' (Beijing) after 1 iteration with α varying from 0 to 1.0 stepped by 0.2. When $\alpha \in [0, +\infty)$, the smaller the parameter α is, the weaker the relevance between the resource allocation and the click frequency will be. If $\alpha = 0$, the resource will be allocated equally. As α grows, the variance of distributed resources increases. When $\alpha = 1.0$, the resource begins to be allocated totally according to the click frequency.

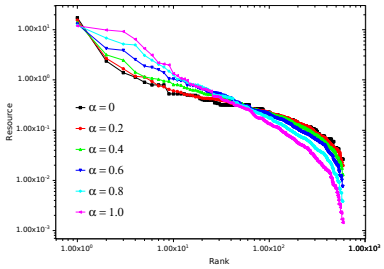

Fig. 6: Log-log plots of the resource distribution of query '北京' (Beijing).

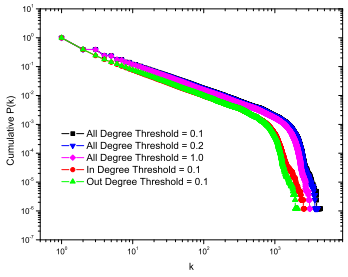

Fig. 7: Log-log plots of cumulative degree distributions of query networks.

Table 4: Properties of query semantic network under threshold $\lambda = 0.1$.

Property	Value
Node Number	834,107
Arc Number	4,735,880
Average All Degree	11.355
Average In/Out Degree	5.678
Average Path Length (Directed)	7.609
Average Path Length (Undirected)	7.231
Clustering Coefficient	0.527
Component Number	556,900
γ	0.915/7.867

Table 5: Examples of the paths on related queries.

Paths on Related Queries
巩俐 → 张艺谋 → 章子怡 → 艺妓回忆录
雅虎 → Yahoo → www.yahoo.com.cn → 雅虎中国
百事可乐 → 可乐 → 可口可乐 → www.icoke.cn → icoke

5 Semantic Networks of Queries

It is straightforward to build query semantic networks via BNRA efficiently. Through the semantic networks we can get much more information among queries. We run one iteration of BNRA for each query and construct a directed and weighted query semantic network by connecting each query to its related queries with threshold $\lambda = 0.1$ which is to discard the related queries with the allocated resource $f_i < 0.1$.

Some properties of the network constructed under threshold $\lambda = 0.1$ are shown in Table 4. Fig. 7 shows the cumulative degree distributions of query networks constructed under different thresholds, and all of them follow power law in the rough, namely $P_c(k) \propto k^{-\gamma}$ where $P_c(k)$ is the cumulative degree distribution, and decay into two parts noticeably which indicates the lack of high degree nodes. All the degree distributions stay stable when the threshold varies from 0.1 to 1.0. The networks show definite small world phenomenon indicating shorter average path length and higher clustering coefficient than the random network of the same size, and scale free effect indicating that the degree distribution follows power law [24]. As shown in Table 5, we display some paths on related queries, which indicates the semantic shift in a sense.

6 Conclusion and Future Work

An asymmetric method was proposed for extracting and quantifying query semantic relations based on network resource allocation using user logs which is simple to implement with low computational cost. We investigated properties of BNRA and found the naive method with only one iteration and allocating

resource by click frequency is good enough for relation extraction. The method is not only 'content oblivious', but also can be easily implemented in a paralleled manner. Possible future work includes: 1) the content based method, such as the common substring method used by Baidu and Google, is expected to be combined with link analysis to achieve more improvement; and 2) more rigorous evaluation will be designed by monitoring the real users choices.

Acknowledgements

This work is supported by the National Science Foundation of China under Grant No. 60621062, 60873174 and the National 863 High-Tech Project under Grant No. 2007AA01Z148. We also thank Peng Li, Qiaxia Jiang and Shaohua Teng for coding work and discussion.

References

1. Beeferman, D., Berger, A.: Agglomerative clustering of a search engine query log. In: Proceedings of the sixth ACM SIGKDD international conference on knowledge discovery and data mining. (2000)
2. Baeza-Yates, R., Tiberi, A.: Extracting semantic relations from query logs. In: Proceedings of the 13th ACM SIGKDD international conference on knowledge discovery and data mining. (2007)
3. Wen, J.R., Jian-Yun, N., Hong-Jiang, Z.: Query clustering using user logs. ACM Transactions on Information Systems **20**(1) (2002)
4. Shen, D., Pan, R., Sun, J.T., Pan, J.J., Wu, K., Yin, J., Yang, Q.: Query enrichment for web-query classification. ACM Transactions on Information Systems **24**(3) (2006) 320–352
5. Beitzel, S.M., Jensen, E.C., Lewis, D.D., Chowdhury, A., Frieder, O.: Automatic classification of web queries using very large unlabeled query logs. ACM Transactions on Information Systems **25**(2) (2007) 9
6. Baeza-Yates, R., Hurtado, C., Mendoza, M.: Query recommendation using query logs in search engines. Workshops on current trends in database technology of 9th international conference on extending database technology (2004)
7. Chirita, P.A., Firan, C.S., Nejdl, W.: Personalized query expansion for the web. In: Proceedings of the 30th annual international ACM SIGIR conference on research and development in information retrieval. (2007) 7–14
8. He, X.F., Yan, J., Ma, J.W., Liu, N., Chen, Z.: Query topic detection for reformulation. In: Proceedings of the 16th international conference on World Wide Web. (2007) 1187–1188
9. Zhou, T., Ren, J., Medo, M., Zhang, Y.C.: Bipartite network projection and personal recommendation. Physical Review E **76**(4) (2007)
10. Deshpande, M., Karypis, G.: Item-based top-n recommendation algorithms. ACM Transactions on Information Systems **22**(1) (2004)
11. Liu, Z.Y., Sun, M.S.: Asymmetrical query recommendation method based on bipartite network resource allocation. In: Proceedings of the 17th international conference on World Wide Web, Beijing (2008)

12. Raghavan, V.V., Sever, H.: On the reuse of past optimal queries. In: Proceedings of the 18th annual international ACM SIGIR conference on research and development in information retrieval. (1995) 344–350
13. Fitzpatrick, L., Dent, M.: Automatic feedback using past queries: social searching? In: Proceedings of the 20th annual international ACM SIGIR conference on research and development in information retrieval. (1997) 306–313
14. Baeza-Yates, R., Hurtado, C., Mendoza, M.: Query clustering for boosting web page ranking. Advances in Web Intelligence (2004) 164–175
15. Sahami, M., Heilman, T.D.: A web-based kernel function for measuring the similarity of short text snippets. In: Proceedings of the 15th international conference on World Wide Web. (2006) 377–386
16. Fonseca, B.M., Golgher, P.B., de Moura, E.S., Ziviani, N.: Using association rules to discover search engines related queries. In: Proceedings of the first conference on Latin American Web Congress. (2003) 66–71
17. Broder, A.: A taxonomy of web search. ACM SIGIR Forum **36**(2) (2002) 3–10
18. Kang, I.H., Kim, G.C.: Query type classification for web document retrieval. In: Proceedings of the 26th annual international ACM SIGIR conference on research and development in informaion retrieval. (2003) 64–71
19. Gravano, L., Hatzivassiloglou, V., Lichtenstein, R.: Categorizing web queries according to geographical locality. In: Proceedings of the 12th international conference on information and knowledge management. (2003) 325–333
20. Baeza-Yates, R.: Graphs from search engine queries. Proceedings of the 33rd conference on current trends in theory and practice of computer science (2007) 1–8
21. Zhou, T., Jiang, L.L., Su, R.Q., Zhang, Y.C.: Effect of initial configuration on network-based recommendation. Europhysics Letters **81**(5) (2008) 58004
22. Ross, S.M.: Introduction to Probability Models, Ninth Edition. Academic Press, Inc., Orlando, FL, USA (2006)
23. Kapp, A.V., Tibshirani, R.: Are clusters found in one dataset present in another dataset? Biostatistics **8**(1) (2007) 9–31
24. Newman, M.E.J.: The structure and function of complex networks. SIAM Review **45**(2) (2003) 167–256

Appendix: Translations of Chinese Queries

For Table 1 and 3: 小说(novel), 玄幻小说(fantasy novel), xiaoshuo(Chinese Pinyin of 'novel'), 言情小说(romantic fiction), 小說(traditional Chinese of 'novel'), 免费小说(novels for free), 言情(romance), 小说阅读网/中文小说网/起点/小说阅读网/潇湘书院/起点中文网/小说网/幻剑书盟/起点中文(names of some Chinese novel websites).

For Table 2: 玄幻小说(fantasy novel), 成人小说(adult fiction), 成人文学(adult literature), 小说(novel), 原创文学(original literature), 文学小说(literature and novel), 读书(reading), 文学家(writers), 艳情文学(erotic literature), 榕树下/世纪文学/天地文学/文学屋/天翼文学/吾爱文学网/起点文学/晋江文学网/晋江文学/星辰变世纪文学/79文学网/天地文学/ 文学城/极品家丁世纪文学/文学殿堂/文学屋(names of some Chinese literature websites).

For Table 5: 巩俐(Gong Li, a Chinese famous actress cooperated with Zhang Yimou), 张艺谋(Zhang Yimou, a Chinese famous director), 章子怡(Zhang Ziyi, a Chinese famous actress cooperated with Zhang Yimou), 艺妓回忆录(Memoirs of a Geisha, a movie starring Zhang Ziyi), 雅虎(Yahoo!), 雅虎中国(Yahoo! China), 百事可乐(Pepsi), 可乐(Cola), 可口可乐(Coca Cola).