Using English Information in Non-English Web Search

Wei Gao*
The Chinese University of Hong Kong
Shatin, N.T., Hong Kong, China
wgao@se.cuhk.edu.hk

John Blitzer*
Computer Science, UC Berkeley
Berkeley, CA 94720-1776, USA
john@blitzer.com

Ming Zhou
Microsoft Research Asia
Beijing 100190, China
mingzhou@microsoft.com

ABSTRACT

The leading web search engines have spent a decade building highly specialized ranking functions for English web pages. One of the reasons these ranking functions are effective is that they are designed around features such as PageRank, automatic query and domain taxonomies, and click-through information, etc. Unfortunately, many of these features are absent or altered in other languages. In this work, we show how to exploit these English features for a subset of Chinese queries which we call linguistically non-local (LNL). LNL Chinese queries have a minimally ambiguous English translation which also functions as a good English query. We first show how to identify pairs of Chinese LNL queries and their English counterparts from Chinese and English query logs. Then we show how to effectively exploit these pairs to improve Chinese relevance ranking. Our improved relevance ranker proceeds by (1) translating a query into English, (2) computing a cross-lingual relational graph between the Chinese and English documents, and (3) employing the relational ranking method of Qin et al. [15] to rank the Chinese documents. Our technique gives consistent improvements over a state-of-the-art Chinese mono-lingual ranker on web search data from the Microsoft Live China search engine.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval.

General Terms

Algorithms, Performance, Experimentation.

Keywords

Learning-to-rank, non-English web search, query translation, cross-lingual similarity metrics.

1. INTRODUCTION

The English web is larger and has existed longer than the web in any other language. Because of this, English search engines have been tuned for longer and with more effort than search engines in other languages. Static link analysis such as PageRank [2], click-through information [10], and query and document classification [11] are all important features for modern web search rankers. These features often don't translate directly to new languages. Because of lack of exposure, useful and important non-English sites often have low PageRank. When entering a new linguistic market, a search engine does not have a large user base, and click-through information can be unreliable. Large, reliable training sets for query and webpage classification are often unavailable in non-English languages.

At the same time, a significant portion of non-English queries have unambiguous translations into English which also function as good English queries. We designate these queries as linguistically non-local (LNL). As an example, the Chinese query "哈利波特" can be translated as "Harry Potter", a query for which abundant and useful English information exists (e.g., sites devoted to the books and movies). For this query, the features from the English documents may be able to provide us with useful information in Chinese. By contrast, the Chinese query "北方人才网" (Northern [China] skilled person network) has no immediate English counterpart. Even if we were able to accurately translate this query into English, the resulting English documents are sparse and not useful.

This work describes a method for improving search quality for linguistically non-local Chinese queries by exploiting English information. Our method falls under the framework of machine learning for search ranking [3, 4, 5, 7, 10, 15, 21]. We train our model using a list of Chinese LNL queries, together with relevance judgments for a list of Chinese documents. Training proceeds as follows: For each Chinese query, we first translate it into English and retrieve a list of English documents [6, 8]. We then use a dictionary-based translation system to compute cross-lingual similarities among the Chinese and English documents [13]. Finally, we use these cross-lingual similarities to learn a relational ranking function for the Chinese documents [15]. We show consistent improvement in relevance ranking, as measured by normalized discounted cumulative gain (NDCG) [9] on a corpus of web search data and relevance judgments from the MSN Live search engine [23, 24].

While our simple procedure does lead to improved search results, we emphasize that this is a first step, and we have by no means exhausted the possibilities for using cross-lingual information to improve search results. Perhaps most strikingly, our current method does not exploit English relevance ranking labels at all. But because English relevance ranking data is also larger and

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
iNEWS'08, October 30, 2008, Napa Valley California, USA.
Copyright 2008 ACM 978-1-60558-253-5/08/10...$5.00.

* This work was done while the authors were visiting Microsoft Research Asia. Both authors contributed to the work equally.

Table 1. Examples of linguistically non-local (left columns) and local (right columns) Chinese queries, together with their English translations or glosses. For the linguistically non-local queries, English results may help us perform better Chinese ranking. For local queries, English results are unlikely to be helpful.

Linguistically non-local Chinese query	English translation	Local Chinese query	English gloss
福特汽车	Ford Motor Company	李白写的诗	The poems of Li-Bai
公共关系	public relations	四川长虹手机	Sichuan Changhong cell phones
哈利波特	Harry Potter	大红鹰	Great Red Eagle [Tobacco]
音乐欣赏	music appreciation	北方人才网	Northern [China] skilled person network

better constructed than other languages, this is an immediate area for further exploration. The latter part of this paper is devoted to exploring current and future approaches for using English information.

The rest of this paper is organized as follows: Section 2 introduces the concept of LNL queries; Section 3 gives a real-world LNL query example that motivates our ranking scheme; in Section 4, we present our ranking model by using relational relevance information across different languages; Section 5 discusses experiments and results; Section 6 presents the related work; and we conclude with a brief discussion of future work in Section 7 and 8.

2. LINGUISTICALLY NON-LOCAL QUERIES

Defining what makes a query linguistically non-local is a difficult problem. Because of this, we use an automatic definition derived from query logs and a large bilingual dictionary. We designate a Chinese query as linguistically non-local if its translation also occurs in the English query log. Table 1 gives several examples of linguistically local and non-local Chinese queries from the query logs of a major search engine. Even if we could translate the linguistically local Chinese queries[1], we cannot expect a large amount of rich English information (when compared to the Chinese). On the other hand, because the translations of the linguistically non-local queries occur in the English query log itself, we know a priori that they yield reasonable queries.

Even if we were able to achieve improvements on LNL queries, it would only be worthwhile if there were a significant number of them to begin with. We selected 32,730 Chinese queries and translated them into English. After automatic translation with a Chinese-English dictionary with 940,000 unique entries, we were left with 7,008 queries whose translations also appeared in an English query log of size about 7.2 million. After manually checking these queries, we found 3,767 that were perfect translations. The final ratio (3,767 / 32,730) yields the estimate that 11.5% of queries are LNL queries. We emphasize, though, that with an improved dictionary and larger query logs, this ratio may rise even higher.

3. A MOTIVATING EXAMPLE

With our taxonomy of Chinese queries as linguistically local and non-local in Section 2, we expect that the relevance ranking of Chinese LNL queries, such as foreign names, globally hot topics, general concepts, etc., will benefit from information contained in the returned documents of their corresponding English queries. For example, given the Chinese query “哈利波特” (Harry Potter), search results from the English query tend to be more relevant than just searching by Chinese because the concept is more popular in the English-speaking world.

Figure 1 shows ranking judgments for Chinese websites retrieved when given the query “哈利波特”. For each retrieved document, we give a human judgment (in bold) ranging from “Bad” to “Excellent”. These judgments were made independently of the English translations and this work. Similarly, if we retrieve English documents using the translation “Harry Potter”, we obtain the results shown in the second column. Some documents among these results are conceptually quite similar, such as C2, E1 & E2 (all are official sites of the movie), C1 & E3 (they are about the “Sorcerer's Stone” story), and C4 & E4 (unofficial fans sites).

If we ignore inter-document similarity and use a ranker based only on the Chinese queries, we obtain the results given in the third column, where the unofficial site http://club.52harrypotter.com is ranked the highest. However, by exploiting cross-lingual similarities, we can use the fact that C1, C2, and C3 are similar to English documents which have strong PageRank, click-through, and domain recognition. From a learning-to-rank perspective, this information can be exploited to help us learn a better ranking function which increases the scores of C1, C2, and C3. Because C4 is not as similar to an official English site, its score will not be increased. Indeed, we chose this example to showcase our improved relevance ranker (shown at the rightmost column). By training a ranker which exploits document similarities, we are able to rank the websites C1, C2, and C3 above C4. The next section describes in detail how we train this ranker to improve the web search ranking for LNL Chinese queries.

[1] The linguistically local queries here did not appear in our dictionary. We provided the glosses ourselves.

Chinese Gold Standard for the query “哈利波特”	English Documents for the query “Harry Potter”	No similarity	With similarity
C1 - http://ent.sina.com.cn/m/f/f/potter1.html **(Good)** 《哈利-波特与魔法石》_影音娱乐_新浪网-片名：Harry Potter and the Sorcerer's Stone 译名：哈利・波特与魔法石/哈利・波特 1 导演：导演克里斯-哥伦布 Chris Columbus 原著：J.K.罗琳 ...	E1 - http://harrypotter.warnerbros.com/main/homepage/intro.html **(Excellent)** Harry Potter - The Official Site The Official Harry Potter Website offers content, games and activities which seamlessly extend the magical world of Harry Potter beyond the big screen…	C4	C1
C2 - http://harrypotter.tw.warnerbros.com/main/homepage/home.html **(Good)** 正式的哈利波特網站- 正式的哈利波特網站來了! 電影預告片，電影片段，霍格華茲的拍片現場，華納兄弟出品的電影哈利波特，神秘的魔法石將活生生地展現 JK 羅琳筆下的巫師和女巫，哈利波特、榮 ...	E2 - http://harrypotter.warnerbros.co.uk/diversions/index.html **(Good)** Harry Potter | Fun & Games The official site of Harry Potter! Movie trailers, film clips, behind the scenes at Hogwarts. JK Rowlings' wizards and witche’s Harry Potter, Ron Weasley, ...	C1	C3
C3 - http://www.52harrypotter.com **(Good)** 哈利波特 52Harrypotter.Com 我爱哈利波特网 新闻中心 | 同人小说 | 下载中心 | 魔法宝典 | 图库中心 | 哈利维基 | 电子期刊 | 反译联盟 | 魔法链 |丽痕书店 | 俱乐部 | 哈利热潮 哈利小说 哈利电影 哈利游戏 哈利产品 哈迷前线 哈利中国 评论反思 魔法妈妈 魔幻世界 魔幻...	E3 - http://us.imdb.com/title/tt0241527/ **(Fair)** Harry Potter and the Sorcerer's Stone (2001) - Plot summaryHarry Potter and the Sorcerer's Stone on IMDb: Movies, TV, Celebs, and more...	C3	C2
C4 - http://club.52harrypotter.com **(Bad)** 欢迎访问哈利迷俱乐部[哈利迷俱乐部] -- Powered By Dvbbs.net,20..最近没有论坛活动今日:28 帖| 昨日:1114 帖 | 最高日:3528 帖主题:6677 | 帖子:228683 | 会员:228338 | 新会员走吗喂狗 -=> 欢迎访问 哈利迷俱乐部 最新创建圈子最活跃圈子最热门圈子 52 哈利社四川分社 (创始人:冷月清霜,...	E4 - http://www.alivans.com/ **(Fair)** Magic Wands for Harry Potter wands fans - See our Magic WandsMagic Wands from Alivan's are handcrafted to meet the requirements of even the great wizard Harry Potter's wand and are similar to Harry Potter wands…	C2	C4
C5 - http://harrypotter.tw.warnerbros.com **(Bad)** Harry Potter and the Order of the Phoenix2007 Warner Bros. Ent. Harry Potter Publishing Rights © J.K.R. Harry Potter characters, names and related indicia are trademarks of and © Warner Bros. ...	E5 - http://news.bbc.co.uk/cbbcnews/hi/specials/harry_potter/ **(Bad)** CBBC Newsround | Specials | Harry PotterCBBC Newsround - Your stories, your world - first! …	C5	C5

Figure 1. Improving the search results of Chinese LNL query “哈利波特” (left-most column) by leveraging the inter-document similarity across different languages, i.e., the relationship with the information in search results of English query “Harry Potter” (second column). Enhanced ranking results can be observed in the right-most column compared to the third column.

4. A RANKING MODEL FOR LNL QUERIES

Given a set of linguistically non-local Chinese queries together with their unranked Chinese documents, we train a ranking model in three steps. First we translate each query into English and use an English search engine to obtain English documents. Then we construct a similarity graph, where nodes represent Chinese and English documents, and edges between nodes represent cross-lingual similarity. Finally we use this graph to train a relational ranking SVM [15]. This procedure is described formally in Figure 2, and the rest of this section is devoted to describing it in detail.

4.1 Query Translation

For each Chinese query, our first step is to identify a corresponding English query. Our query translation uses a large static dictionary based on a statistical query translation model [6], and we do not investigate the quality of our query translation in this work. In our experiments, we post-process our training and testing data manually to obtain Chinese-English query pairs that we are certain are correct. In deploying a real ranker, of course, we would not have the option of manually post-processing queries to ensure that they are correct valid LNL query pairs. But we emphasize that there has been significant research in the area of bilingual lexicon extraction [8], and we expect that our dictionary can be significantly improved.

4.2 Cross-lingual Similarity

Once we have obtained an English query and corresponding list of documents, the crucial next step (step 2 in Figure 2) is to determine the similarity between Chinese and English documents. We define a similarity score $sim(c,e)$ between a Chinese and English document to be a function mapping pairs of documents to a positive real number. Intuitively a good similarity measure is one which maps cross-lingual relevant documents together, and maintains a large distance between irrelevant Chinese documents and relevant English documents and vice-versa.

We use the similarity measure proposed by Mattieu et al. [13]. Using the same dictionary as for query translation, we let $T(c,e)$ indicate the set of pairs (w_c, w_e) such that w_c is a word in Chinese document c, w_e is a word in English document e, and w_e is the English translation of w_c. We define $tf(w_c, c)$ and $tf(w_e, e)$ to be the term frequency of term w_c in document c and w_e in document e, respectively. Let $df(w_c)$ be the Chinese document frequency for term w_c (with an analogous English definition). If n_c is the total number of Chinese documents, then

$$idf(w_c) = \log\frac{n_c}{df(w_c)}.$$

```
Input:  Chinese queries and documents
$\{q_i,\{c_{ij}\}_{j=1}^{m_{ci}}\}_{i=1}^{n}$
Output: Learned ranking model which maps
from query-document pairs to real-valued
scores $f:(q,c)\to\Re$
(1)  For each query $q_i$
        Translate $q_i$ into English to
        obtain English documents $\{e_{ik}\}_{k=1}^{m_{ei}}$
(2)  For each query $q_i$
        Using a bi-lingual dictionary,
        compute a similarity adjacency
        matrix, with $\mathbf{R}_i(j,k)=sim(c_{ij},e_{ik})$
        Let $\mathbf{L}_i=\mathbf{D}_i-\mathbf{R}_i$ be the graph
        Laplacian for $\mathbf{R}_i$
(3)  Let $\lambda$, $\beta$ be free parameters and
     $Y_{jk}^i=\begin{cases}1, & rank(c_{ij})>rank(c_{ik})\\ \text{-}1, & rank(c_{ij})<rank(c_{ik})\\ 0, & rank(c_{ij})=rank(c_{ik})\end{cases}$ be the entry
     of pair-wise label constraint matrix
     for query $q_i$. Define $\hat{\mathbf{C}}_i=\mathbf{C}_i[(\mathbf{I}+\beta\mathbf{L}_i)^{-1}]^T$
     Return the solution to the
     minimization problem
     $\min_{f}\|f\|^2+\lambda\sum_{i=1}^{n}\sum_{j=1}^{m_{ci}}\sum_{\substack{k=j+1,\\ Y_{jk}^i\neq 0}}^{m_{ci}}\max\left[Y_{jk}^i\,f^T(\hat{c}_{ij}-\hat{c}_{ik})+1,0\right]$
```

Figure 2. Our algorithm for training a ranker for linguistically non-local Chinese queries based on RRSVM.

Bilingual idf is defined as

$$idf(w_c,w_e)=\log\frac{n_c+n_e}{df(w_c)+df(w_e)}.$$

If letting $\overline{T}(c,e)$ denote the set of terms in c that have no translation in e and likewise $\overline{T}(e,c)$ denote the set of terms in e that have no translation in c, we can define the similarity between two documents as

$$sim(c,e)=\frac{\sum_{(w_c,w_e)\in T(c,e)}tf(w_c,c)tf(w_e,e)idf(w_c,w_e)^2}{\sqrt{Z}}$$

where

$$Z=\left[\sum_{(w_c,w_e)\in T(c,e)}\left(tf(w_c,c)idf(w_c,w_e)\right)^2+\sum_{w_c\in\overline{T}(c,e)}\left(tf(w_c,c)idf(w_c)\right)^2\right]\times\left[\sum_{(w_c,w_e)\in T(c,e)}\left(tf(w_e,e)idf(w_c,w_e)\right)^2+\sum_{w_e\in\overline{T}(e,c)}\left(tf(w_e,e)idf(w_e)\right)^2\right].$$

This similarity function can be understood as a cross-lingual analog to the commonly used mono-lingual cosine similarity function.

4.3 Relational Relevance Ranking

Once we know the most similar English documents for a particular Chinese document, we need to use these similarities to help us learn a better ranking function. The relational ranking support vector machine (RRSVM) [15] is a variant of the ranking support vector machine (RSVM) [7] that includes, in addition to the ranking objective, a constraint which encourages similar documents to have scores that are close to one another. In our case, if a Chinese document is similar to an English document with high PageRank or click-through features, the RRSVM objective will automatically encourage the Chinese document to have a higher score (provided those features are in fact useful).

Let λ and Y_{jk}^i be defined as in Figure 2. Here Y_{jk}^i is the $(j,k)^{\text{th}}$ entry of the constraint matrix for query q_i. Each entry indicates one of (-1, 0, +1) depending on whether Chinese document c_{ij} is less relevant, equally relevant, or more relevant to the query than document c_{ik}. By absorbing the margin constraints into the objective function, we can now write the ranking SVM (RSVM) objective as

$$\min_{f}\|f\|^2+\lambda\sum_{i=1}^{n}\sum_{j=1}^{m_{ci}}\sum_{\substack{k=j+1,\\ Y_{jk}^i\neq 0}}^{m_{ci}}\max\left[Y_{jk}^i\,f^T(\hat{c}_{ij}-\hat{c}_{ik})+1,0\right]$$

where λ is a free regularization parameter.

Now let us construct a weighted bipartite similarity graph for each query, where the edge weights for the graph are given by the similarity score $sim(c,e)$. Let the adjacency matrix be defined as in step (2) of Figure 2. For a fixed document scoring function f and query q_i, the RRSVM finds scores z_j which minimize

$$\sum_j(f^Tc_{ij}-z_j)^2+\frac{\beta}{2}\sum_{j,k}\mathbf{R}_i(j,k)(z_j-z_k)^2$$

This quadratic can be solved in closed form. The solution is the minimum energy harmonic function for the graph characterized by $\mathbf{R}_i(j,k)$ [22]:

$$\mathbf{z}=(\mathbf{I}+\beta(\mathbf{D}_i-\mathbf{R}_i))^{-1}\mathbf{C}_i^Tf$$

where $\mathbf{D}_i(j,j)=\sum_k\mathbf{R}_i(j,k)$ is a diagonal matrix. The matrix $\mathbf{L}_i=\mathbf{D}_i-\mathbf{R}_i$ is the graph Laplacian for the query graph with adjacency matrix $\mathbf{R}_i$. Finally, now that we know the form of the scoring function, we may solve for the optimal f which satisfies it. This yields the optimization problem from step (3) of Figure 2.

Qin et al. [15] compute a scoring function based on ranking constraints for all of the documents corresponding to a particular query. That is, they introduce labeled constraints for every node in the graph. In contrast, we are interested only in the Chinese documents for a particular query. The English documents appear as similarity nodes in the graph, and thus in the transformation $(\mathbf{I}+\beta(\mathbf{D}_i-\mathbf{R}_i))^{-1}$, but they don't appear in the final optimization objective. Similarly at test time, we are only interested in the English documents insomuch as they influence the scores of the Chinese documents.

In Sections 2 and 3, we mentioned that our goal is to make use of English features such as PageRank and click-through that may be unavailable or unreliable in other languages. In our technique, these features appear in the transformed instance matrix $\hat{\mathbf{C}}_i = \mathbf{C}_i[(\mathbf{I} + \beta\mathbf{L}_i)^{-1}]^T$ but not in the original matrix $\mathbf{C}_i$. Unfortunately, because of the non-linear matrix inversion, this is difficult to express directly in terms of the original feature space. In general for a particular Chinese document, however, its transformed version is "smoothed" to look more like nearby English (and indirectly, Chinese) documents.

5. EXPERIMENTS AND RESULTS

In this section we give the results of a series of experiments using our proposed algorithm in web search ranking for LNL queries.

5.1 Dataset and Baselines

Because this workshop focuses on web search, all of our experiments are performed on annotated Chinese and English data of the MSN Live search engine [23, 24]. As we mentioned in Section 2, we built our training set by choosing randomly a subset of (labeled) Chinese queries from the Chinese query log and automatically translating them into English. After this, we manually choose 803 pairs of queries which are accurate translations. The average number of annotated Chinese documents for each of these queries is 25, but there is a large amount of variability (The minimum number of documents is 5 and the maximum number is 50). We end up with about 7,000 Chinese and 10,000 English documents in the data set, respectively. Each document is annotated from 0 (irrelevant) to 5 (perfect).

For each web page of a given query, the features consist of query-dependent features (e.g., term frequency) and query-independent features (e.g., PageRank) extracted from the page and the index. There are 352 such features in total.

Our baseline is a ranking SVM, trained only on Chinese documents. All of the results we report here are 4-fold cross-validated, with (approximately) 600 queries being used as training and 200 as testing. All of our results are trained using stochastic gradient descent [17] on the loss functions of the RSVM (see Section 4.3) and the RRSVM (see step (3) in Figure 2).

5.2 Mono-lingual and Joint Similarities

In Section 4, we described a bipartite graph based on the cross-lingual similarity (see Section 4.2). However, it may be that mono-lingual similarities alone can give improvement in ranking accuracy, or that a graph with both cross-lingual and mono-lingual edges can be more effective than a graph with only mono-lingual edges. We define the mono-lingual similarity between two documents to be the $(tf \times idf)$ – weighted cosine similarity. Throughout the rest of this section, mono-lingual similarity refers to a graph built only from Chinese information (and ignoring English). Cross-lingual similarity refers to bipartite graphs of the type described in Section 4, where there are Chinese-English edges, but no Chinese-Chinese edges. Joint similarity graphs are those built with both mono-lingual and cross-lingual edges.

As Section 4 indicates, there are several hyper-parameters that can be tweaked in model design. We investigate varying some of these hyper-parameters in Section 5.3. In the end, though, our goal is to compare ranking with cross-lingual and mono-lingual information. In order to do this, we select a single ranker for each of our mono-lingual, cross-lingual, and joint similarity graphs and show normalized discounted cumulative gain (NDCG) [9] for these models (see Table 2. NDCG is an IR evaluation metric that can handle multiple levels of relevance following the principles that highly relevant documents are more valuable than marginally relevant ones, and the document with a higher ranking position is more valuable because it is more likely to be examined by the user than that with a lower ranking position). Then we selected each of these results by choosing the best possible settings of hyper-parameters k (graph local neighborhood size) and β . The optimal hyper-parameters are determined by maximizing the average relative gain in NDCG. That is, for each of mono-lingual, cross-lingual and joint similarity graphs, we choose the best setting of k and β for the average relative improvement in NDCG over the no-graph baseline at different ranking positions. For a particular model m , we write $NDCG_m$ to be the model NDCG and $NDCG_{ng}$ to be the NDCG of the "no-graph" model, which ignores similarity information. Then the average relative improvement in NDCG (over ranking positions 1, 3, and 5 in our case) is defined as

$$\frac{1}{3}\frac{NDCG_m@1 - NDCG_{ng}@1}{1 - NDCG_{ng}@1} + \frac{1}{3}\frac{NDCG_m@3 - NDCG_{ng}@3}{1 - NDCG_{ng}@3} + \frac{1}{3}\frac{NDCG_m@5 - NDCG_{ng}@5}{1 - NDCG_{ng}@5}$$

Table 2. Performance of RRSVM in terms of NDCG@1,3,5 examined under different similarity measures and compared with the RSVM baseline without using the similarity graph. The optimal parameters are resolved by maximizing the average relative gains over NDCG@1,3,5

	NDCG@1	NDCG@3	NDCG@5
no-graph (baseline)	65.61	74.08	78.38
mono-lingual (k=5, β=0.4)	66.38 (+1.17%)	74.78 (+0.94%)	78.62 (+0.31%)
cross-lingual (k =20, β =0.2)	66.74 (+1.72%)	**74.81** **(+0.99%)**	78.94 (+0.71%)
joint (k-all, β =0.5)	**66.9** **(+1.97%)**	74.46 (+0.51%)	**78.97** **(+0.75%)**

As we can see, the RRSVM consistently outperforms the no-graph baseline, which implies the effectiveness of using a similarity graph. Furthermore, at these settings using cross-lingual similarity improves over mono-lingual similarity for NDCG@1, 3, and 5. This suggests that while not perfect, the cross-lingual similarity measure does capture useful similarity information among the documents across different languages.

It is also worth noticing that the RRSVM improves NDCG@1 much better than the performance at the rest of the positions. The mono-lingual, cross-lingual and joint similarities can boost NDCG@1 above the baseline by 1.17%, 1.72% and 1.97% respectively, while the improvements at 3 and 5 are clearly less than 1%, and also NDCG@3 is basically improved larger than NDCG@5 except for using joint similarity. We don't have a good explanation for why joint similarity performs comparably worse at

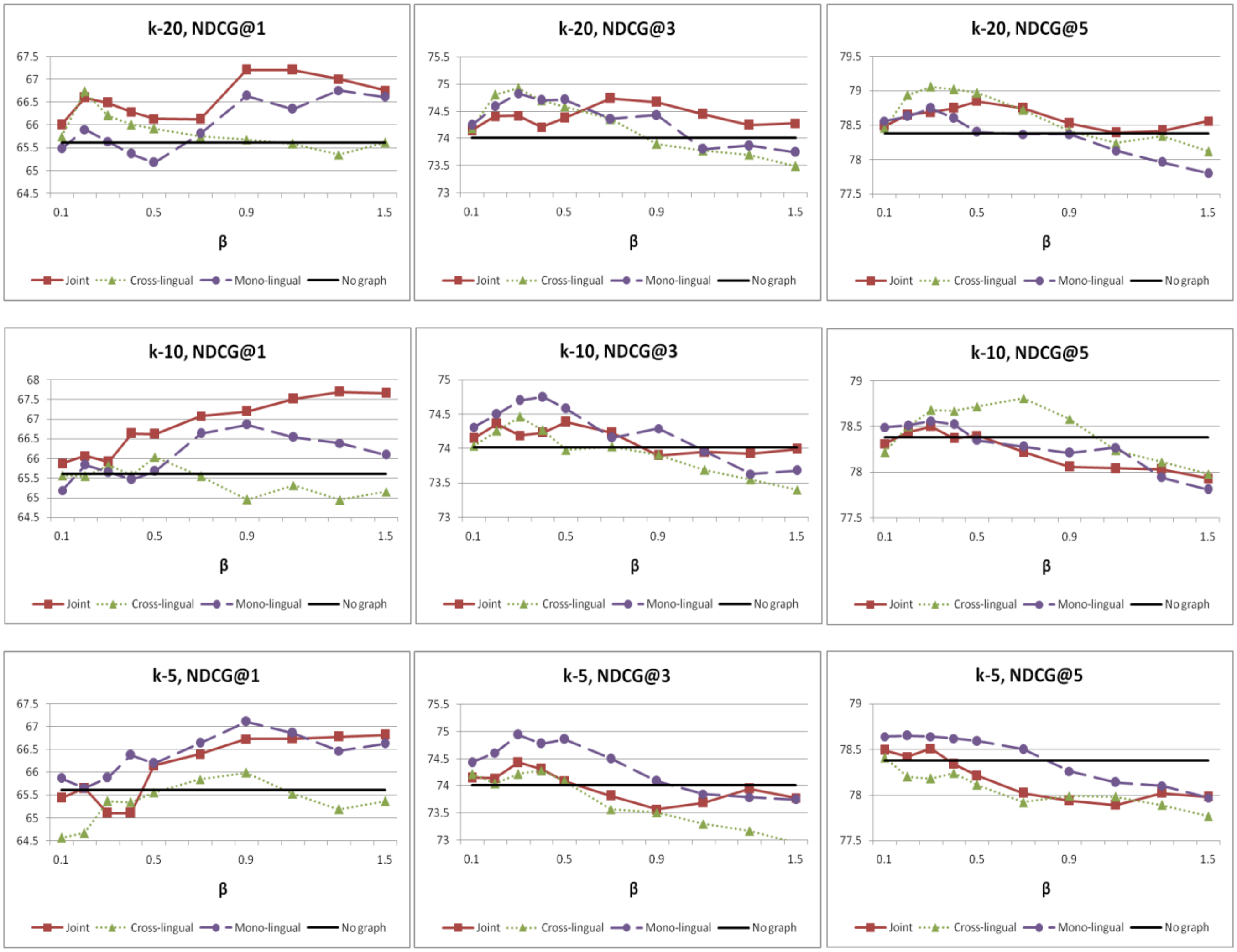

Figure 3. The comparison of different similarity measures contributing to the performance of Chinese documents ranking in terms of NDCG@1,3,5. The number of nearest neighbors *k* and the tradeoff parameter *β* are varied to show their influences.

NDCG@3, and ultimately we don't have a complete understanding of how cross-lingual and mono-lingual similarities interact. This is an important area for further investigation.

5.3 Graph Topology and β

It is often the case that in high dimensional vector spaces such as ours, local similarities can be more reliable than similarities among more distant points. Because of this, we also investigate truncating the edges for each node to its nearest neighbors. Finally, we examine how the distance changes as we increase the tradeoff parameter β (see Figure 3), which governs the relative tradeoff between the similarity graph and the ranking function. When $\beta = 0$, we use only the ranking function, and as we increase β, we increasingly penalize documents which are nearby in the graph but have dissimilar scores.

Figure 3 illustrates varying neighborhood size k (for *k*=5, 10 and 20 nodes in the graph) and β (on the x-axis). As a general trend, we observe that all of the similarities perform well under larger k. In particular, the joint similarity is consistently above the no-graph baseline when $k = 20$. Its performance degrades with fewer numbers of nearest neighbors used for the cases of NDCG@3 and 5. Cross-lingual similarity performs well under similar situations, but is less robust to smaller values of k. However, cross-lingual similarity outperforms both other methods for large k and small β, resulting in the largest overall average relative gain in NDCG. Finally, mono-lingual similarity performs best when $k = 5$.

The variance with k seems to indicate that mono-lingual similarity is accurate on a per-document basis, but overall it has limited ability to improve ranking. In contrast, our current cross-lingual similarity measure is accurate only in aggregate across many documents. Developing a way to improve cross-lingual similarity or to interpolate more accurately between cross-lingual, joint, and mono-lingual similarities is a topic for further research (see section 7). Finally, we note that while the variation of NDCG in β is fairly smooth, the graphs do not show a convex shape with a single clear maximum. Indeed, NDCG may decrease with

β before increasing again. Unfortunately, this is due to the complex nature of the inverse Laplacian and the NDCG measure itself, both of which are non-linear functions. We also plan a more thorough investigation of exactly how NDCG and other evaluation metrics interact with our hyper-parameters.

5.4 Illustrative Examples

Here we will give several illustrative examples using the queries and the ranking results from our dataset.

For example, given the query "皇家马德里" (Real Madrid), we have the Chinese website of Real Madrid Fans Club (皇家马德里球迷俱乐部|皇马中文网站，http://www.realmadridfans.com) ranked at the 6-th position by RSVM. According to our human judgments, this page should be ranked third. Using our cross-lingual similarities, we promote the site to the 4th position by RRSVM. This is because the English homepage of the "Real Madrid Fan Community" (http://www.realmadrid.dk/), labeled as "excellent", has high similarity with this Chinese page.

Another interesting example is from the Chinese query "丰田" (Toyota). Although the RSVM baseline ranks the Chinese homepage of Toyota (http://www.toyota.com.cn) as the 6-th position, RRSVM can promote it to the third. In this case, we use both mono-lingual and cross-lingual similarity information. First, the Chinese page for "Vios" (威驰，http://www.vios.com.cn), a popular automobile product of a joint-venture firm with Toyota in China, is an "excellent" result for the query and is highly similar to http://www.toyota.com.cn. Cross-lingually, Toyota's English homepage is also a nearest neighbor of the Chinese homepage.

6. RELATED WORK

Learning to rank is a broad, and in recent years very popular field. Our work does not address different mechanisms for supervised learning of ranking functions, and we cannot possibly hope to cover this entire area in detail here. We briefly mention that although classification and metric regression [4] can be used to model ranking functions, the most widely-used methods typically attempt to model a pair-wise ordering loss among documents [3, 5, 7, 10, 21]. That is, given a particular query q_i, for each pair of documents d_{ij} and d_{ik} such that d_{ij} is more relevant than d_{ik}, the loss is some function of the difference in scores between the two documents. The RSVM model [7] which we use as our baseline falls into this category. More recent methods have also investigated directly optimizing IR evaluation measures [21]. While using these methods could potentially improve our results, we decided against them for the sake of simplicity. None of these methods consider inter-document similarity as useful information. As we mentioned before, the relation ranking SVM of Qin et al. [15] considers inter-document similarity, but they consider only mono-lingual similarity among the English documents.

The other closely-related area is multi-lingual information retrieval (MLIR). The goal of MLIR systems is to simultaneously rank documents in multiple languages. In contrast, we focus on using multi-lingual information to rank documents in a *single* language. Because of this difference in goals, most existing work in MLIR focuses on heuristics for merging ranked lists from multiple languages [1, 16, 18]. While it is not our primary focus, it is possible that the techniques we outline in this paper would also be useful for multi-lingual ranking.

7. FUTURE WORK

We believe our results in Section 5 demonstrate that there is useful information available from English rankers for Chinese data. But our current algorithm has by no means exhausted the possibilities for exploiting cross-lingual information in mono-lingual search.

7.1 Improved Similarity

Perhaps the most obvious improvement is a better cross-lingual (and mono-lingual) similarity measure. Our current similarity measures do not use state-of-the-art bi-lexicon mining techniques or machine translation, and they make no attempt to distinguish text in different sections of web pages or images. At the same time, it seems natural to assume that an effective similarity score would involve some combination of page layout, varying levels of text understand and machine translation, and perhaps even image understanding. With such a large number of factors, it makes sense to consider learning a similarity function. While there has been work on learning the parameters of random walks [20], which are closely related to the Gaussian random field methods of the RRSVM [22], it is unclear if these could be directly applied. This remains an important topic for further research.

7.2 More Sources of Cross-Lingual Information

When considering the ways English information could be used in non-English ranking, we can construct a rough taxonomy of non-English queries. At a high level, there are 4 categories

(1) Linguistically local queries. The majority of Chinese queries have no corresponding useful English counterpart. But that does not mean that the information available to an English ranker is completely useless cross-lingually. It could be that some useful information could be transferred from one model to another directly.

(2) LNL queries with Chinese relevance annotated. This is the subject of this work. As we discussed in 7.1, there is still more to be done in this specific case.

(3) LNL queries with only English annotated. For most major search engines, English has many more relevance judgments than other languages. This is true for LNL queries as well. Can we develop models which exploit this information more directly in non-English search?

(4) LNL queries with neither language annotated. The vast majority of LNL queries fall into this category, where there is no direct label feedback. But we do know that since the queries correspond, the rankings should (partially) correspond as well. This is most similar to multi-view semi-supervised learning [14], and cross-lingual multi-view learning has been explored extensively in natural language processing [12, 19].

8. CONCLUSION

While English web pages continue to characterize the majority of the web, non-English search is becoming increasingly important. At the same time, the major search engines all have much better English rankers than non-English rankers. In part, we believe this to be due to the number and reliability of features like PageRank, click-through, and web page and query categorization. This work investigated exploiting English information for a subset of Chinese queries which we called linguistically non-local queries.

We showed how to use query logs and a large bilingual dictionary to improve relevance scores in Chinese for these queries. We emphasize that while our work is encouraging, we believe we have only scratched the surface of the amount of ways we can potentially exploit cross-lingual information for non-English ranking.

9. ACKNOWLEDGEMENTS

We thank Tie-Yan Liu, Cheng Niu, and Tao Qin for their helpful advice throughout the course of this work. Tao Qin also provided us with a fast implementation of RRSVM, which we used in the experiments of this work.

10. REFERENCES

[1] Braschler, M. and Peters, C. Cross-Language Evaluation Forum: Objectives, Results, and Achievements. *Information Retrieval*, 7(1-2):7-31, 2004.

[2] Brin, S. and Page, L. The Anatomy of a Large-Scale Hypertextual Web Search Engine. In *Proc. WWW 1998*.

[3] Burges, C., Shaked, T., Renshaw, E., Lazier, A., Deeds, M., Hamilton, N. and Hullender, G. Learning to Rank Using Gradient Descent. In *Proc. ICML 2005*, pages 89-96.

[4] Crammer, K. and Singer, Y. PRanking with Ranking. In *Proc. NIPS 2002*.

[5] Freund, Y., Iyer, R., Schapire, R. and Singer, Y. An Efficient Boosting Algorithm for Combining Preferences. *Journal of Machine Learning Research*, 4:933-969, 2004.

[6] Gao, J. F., Nie, J.-Y., Xun, E., Zhang, J., Zhou, M., and Huang, C. Improving Query Translation for CLIR Using Statistical Models. In *Proc. ACM SIGIR 2001*, pages 96-104.

[7] Herbrich, R., Graepel, T. and Obermayer, K. Support Vector Learning for Ordinal Regression. In *Proc. ICANN 2003*, pages 97-102.

[8] Huang, F., Zhang, Y., and Vogel, S. Mining Key Phrase Translations from Web Corpora. In *Proc. EMNLP 2005*, pages 483-490.

[9] Jarvelin, K. and Kekanainen, J. IR Evaluation Methods for Retrieving Highly Relevant Documents. In *Proc. ACM SIGIR 2000*, pages 41-48.

[10] Joachims, T. Optimizing Search Engines Using Clickthrough Data. In *Proc. ACM SIGKDD 2002*, pages 133-142.

[11] Kang, I. H. and Kim, G. C. Query Type Classification for Web Document Retrieval. In *Proc. ACM SIGIR 2003*, pages 64-71.

[12] Klementiev, A. and Roth, D. Weakly Supervised Named Entity Transliteration and Discovery from Multilingual Comparable Corpora. In *Proc. ACL 2006*, pages 817-824.

[13] Mattieu, B. Besancon, R., and Fluhr, C. Multilingual Document Clusters Discovery. In *Proc. Recherche d'Information Assistée par Ordinateur (RIAO) 2004*, pages 1-10.

[14] Mitchell, T. and Blum, A. Combining Labeled and Unlabeled Data with Co-training. In *Proc. Conference on Learning Theory (COLT) 1998*, pages 92-100.

[15] Qin, T**.**, Liu, T. Y., Zhang, X. D., Wang, D. S., Xiong, W. Y., and Li, H. Learning to Rank Relational Objects and Its Application to Web Search. In *Proc. WWW 2008*, pages 407-416.

[16] Savoy, J. and Berger P. Y. Selection and Merging Strategies for Multilingual Information Retrieval. In *Proc. CLEF 2004, LNCS 0302*, 2005.

[17] Shalev-Shwartz, S., Singer, Y., and Srebro, N. Pegasos: Primal Estimated sub-GrAdient SOlver for SVM. In *Proc. ICML* 2007, pages 807-814.

[18] Si, L. and Callan, J. A. Multilingual Retrieval by Combining Multiple Multilingual Ranked Lists. In *Proc. CLEF 2005, LNCS 4022*, 2006.

[19] Snyder, B. and Barzilay, R. Unsupervised Multilingual Learning for Morphological Segmentation. In *Proc. ACL 2008*.

[20] Toutanova, K. Ng, A., and Manning, C. Learning Random Walk Models for Inducing Word Dependency Distributions. In *Proc. ICML 2004*.

[21] Yue, Y., Finley, T., Radlinski, F., and Joachims, T. A Support Vector Method for Optimizing Average Precision. In *Proc. ACM SIGIR 2007*, pages 271-278.

[22] Zhu, X., Ghahramani, Z., and Lafferty, J. Semi-Supervised Learning Using Gaussian Fields and Harmonic Functions. In *Proc. ICML 2003*, pages 912-919.

[23] MSN Live Search. http://live.com.

[24] MSN Live Search Chinese. http://live.com.cn.