XKnowSearch! Exploiting Knowledge Bases for Entity-based Cross-lingual Information Retrieval

Lei Zhang
Karlsruhe Institute of Technology (KIT)
76128 Karlsruhe, Germany
l.zhang@kit.edu

Michael Färber
Karlsruhe Institute of Technology (KIT)
76128 Karlsruhe, Germany
michael.faerber@kit.edu

Achim Rettinger
Karlsruhe Institute of Technology (KIT)
76128 Karlsruhe, Germany
rettinger@kit.edu

ABSTRACT

In recent years, the amount of entities in large knowledge bases available on the Web has been increasing rapidly, making it possible to propose new ways of intelligent information access. Within the context of globalization, there is a clear need for techniques and systems that can enable multilingual and cross-lingual information access. In this paper, we present XKnowSearch!, a novel entity-based system for multilingual and cross-lingual information retrieval, which supports keyword search and also allows users to influence the search process according to their search intents. By leveraging the multilingual knowledge base on the Web, keyword queries and documents can be represented in their semantic forms, which can facilitate query disambiguation and expansion, and can also overcome the language barrier between queries and documents in different languages.

1. INTRODUCTION

The Web has radically altered the way that information is shared by lowering the barrier to publishing and accessing documents. With more than one trillion pages, the Web has become a *global document repository*, which encompasses practically almost every topic of human interest. As the founding language, English has always dominated the Web, where it is estimated that more than half of all Web content is in English. However, the share of English Web pages decreases and that of other languages increases rapidly, which ensures the multilingual viability of the Web.

Accessing Web documents can be efficient when the information needs of users are expressed as keywords. However, both documents and keyword queries are usually treated as plain text by current search engines, i.e., term-based matching algorithms are used to retrieve the relevant documents according to a given information need. Due to the problem for ambiguous terms, there exists a *semantic gap* between unstructured text and its actual meaning.

As a consequence of the ability to understand more than one language, many Web users are interested in relevant information in multiple languages, especially when relevant documents on the Web are scare in their native language. With the goal that users from different countries have access to the same information on the Web, there exists a *language barrier* for cross-lingual access to information originally produced for a different culture and language.

On the other hand, the Semantic Web has come a long way with the goal of extending the existing Web by bringing semantics to its content. There has been an increasing effort in which the Semantic Web community has envisioned how semantics and the Web can be combined. Linked Open Data (LOD)[1] is such a way of publishing semantic data on the Web that gives humans and machines direct access to such structured data [1]. It is important to note that many LOD sources are generally in multiple languages. As an example, DBpedia[2], staying in the center of the LOD cloud, is a crowd-sourced community effort to extract structured knowledge from multilingual Wikipedia, resulting in localized versions of DBpedia in more than 100 languages, and to make this information available on the Web [5].

The ever-increasing quantities of semantic data in large knowledge bases (KBs) on the Web, such as DBpedia, Freebase and YAGO, pose new challenges but at the same time open up new opportunities of intelligent information access. These knowledge bases contain a vast amount of entities and the knowledge about them such that the Web also serves as a *global knowledge repository* of entities. Due to an increasing portion of queries involving entities for Web search [6], the exploitation of *entities and their knowledge* beyond the term-based paradigm for information retrieval (IR) has become an area of particular interest.

In this paper, we present XKnowSearch!, a novel system for multilingual and cross-lingual IR by exploiting entities and their relations in the KB with the goal of addressing the following challenges that traditional keyword search systems mainly suffer from:

Inflexibility. Keyword search has proven to be a simple and intuitive paradigm for expressing the information needs. However, traditional keyword search systems do not allow users to be involved in the search process to perform query refinement according to their search intents.

Ambiguity. Keyword queries are naturally ambiguous due to the fact that keywords could refer to different things in different contexts. This problem is more serious in the multilingual and cross-lingual setting, because the same keywords could have different meanings in different languages.

Incompleteness. Keyword queries are often incomplete in the sense that only the aliases, acronyms and misspellings

CIKM '16, October 24–28, 2016, Indianapolis, IN, USA.

ACM ISBN 978-1-4503-4073-1/16/10..

DOI: http://dx.doi.org/10.1145/2983323.2983324

[1] http://lod-cloud.net/

[2] http://dbpedia.org/

are usually given in the queries. In addition, keyword queries might contain concept names, e.g., "*online companies of US*", which could refer to a set of entities.

Cross-linguality. Multilingual users probably formulate their information needs using native language, but they are interested in relevant information in any language that they can understand. Moreover, specifying the query language should not be the burden of users and they could even issue queries consisting of keywords in multiple languages.

Concerned with these challenges, XKnowSearch! supports keyword search on Web documents by representing the information needs of users as entity graphs in the KB to avoid the semantic ambiguity of keyword queries. Based on various query interpretations, it allows users to explore entity relations in the KB to further refine the queries and also enables automatic query expansion. In order to bridge the language barrier between queries and documents, we leverage the multilingual KB, namely DBpedia, to construct semantic representations of queries and documents in different languages and use them to develop a more effective way of modeling the document relevance on the basis of entity knowledge for satisfying the information needs of users.

2. LIMITATIONS OF EXISTING SYSTEMS

In this section, we review the existing entity-based search systems and discuss their limitations, which serve as the motivation of our XKnowSearch! system.

EntEXPO [4] provides entity-based query expansion by finding a list of related entities of a single query entity and it allows users to manually adjust the weight of each related entity. However, there is no discussion about how to resolve the ambiguity of the query keywords and it does not concern with the queries containing concept names or multiple entities. EntEXPO seems to support search only in English.

Kuphi [2] employs semantic annotations of documents to enhance the performance of document retrieval. It allows interactive query reformulation by selecting the intended entity and adjusting the weights of related entities. However, the system assumes that a keyword query is a single entity name such that it cannot handle queries containing more than one entity name or concept names.

STICS [3] has been proposed to support users in searching for terms, entities and categories. However, users have to specify the query entities and categories explicitly such that the ambiguity of queries can only be resolved by users. Moreover, it supports neither query expansion with related entities nor interactive query formulation / refinement. Finally, STICS also does not support cross-lingual search.

Recently, almost every major commercial Web search engine has announced their work on incorporating entity information from knowledge bases into its search process, including Google's Knowledge Graph, Yahoo!'s Web of Objects and Microsoft's Satori Graph / Bing Snapshots. However, there are still some limitations. Firstly, most search engines take into account only the most prominent entities matching the keyword query. Secondly, they can only understand individual entities, but cannot deal with a set of entities expressed by a concept name. For example, given the keyword query "Internet companies of US", they do not suggest the expected entities, such as Google and Yahoo!, and the retrieved documents are mostly only matched against the query keywords, such as "Internet companies" and "US". Finally, they do not support cross-lingual search. For example, given the Chinese query "马云" denoting Jack_Ma, the

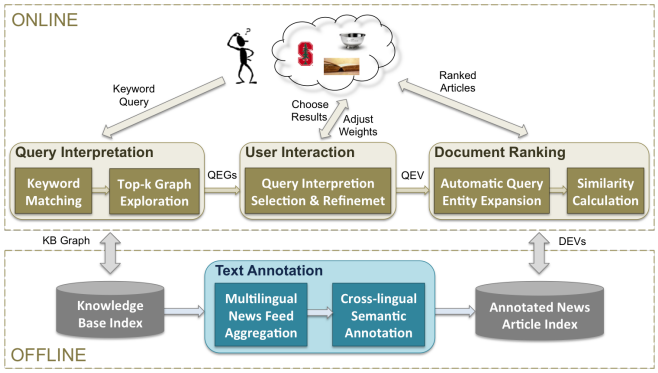

Figure 1: The System Architecture of XKnowSearch!.

founder of Alibaba_Group, they rarely retrieve any relevant English documents about Jack_Ma or Alibaba_Group.

In summary, existing entity-based search systems cannot well address the challenges of *inflexibility*, *ambiguity* and *incompleteness*. More importantly, all of them do not support *cross-lingual search*. For example, EntEXPO seems to support only English and STICS supports both English and German, but neither of them can handle cross-lingual search. Although Kuphi enables users to search documents in one language by using queries in another language, users have to specify the input language of the query, which can only be a single entity name. To the best of our knowledge, XKnowSearch! is the first entity-based system to multilingual and cross-lingual IR with the goal of addressing these challenges, where users can issue keyword queries in any language, which can even contain keywords in multiple languages, for retrieving multilingual documents, especially in any other languages. In order to avoid the users' burden of specifying the query languages, XKnowSearch! does not assume any input language given by users.

3. SYSTEM ARCHITECTURE

The system architecture of XKnowSearch!, as shown in Fig. 1, consists of a set of components. While *text annotation* is performed offline, *query interpretation*, *user interaction* and *document ranking* are handled online. In this section, we briefly introduce these components.

Text Annotation. For offline processing, we first employ a *news feed aggregator*[3] to acquire a multilingual real-time stream of news articles on the Web across the world. The collected articles are in various languages, such as English (50% of all articles), German (10%) and Chinese (5%).

Then *cross-lingual semantic annotation* is performed to enrich the collected news articles in different languages with entities in the KB. More specifically, the KB we use, namely DBpedia, contains a formal representation of entities and semantic relations between them. In addition, it is multilingual, i.e., there are multiple language versions containing entities grounded in different languages. Based on that, we employ our cross-lingual semantic annotation system[4] to annotate the multilingual documents with entities grounded in one or more hub languages [8]. It helps to bridge the ambiguity of natural language text and precise formal semantics captured by the KB as well as to transform documents in different languages into a unified representation. The result-

[3]http://newsfeed.ijs.si
[4]http://km.aifb.kit.edu/sites/xlisa/

ing annotated documents are stored in an inverted index to make them searchable with KB entities.

Query Interpretation. The online process starts with a keyword query in any language (even with keywords in multiple languages). Instead of retrieving documents directly by keywords, XKnowSearch! first finds the *query entity graphs* (*QEGs*) matching the keyword query by exploring the semantic graph of the KB[5] with nodes representing entities and edges describing their relations.

The first step of *query interpretation* is *keyword matching*. To address the challenge of matching query keywords in different languages to entities, we constructed a cross-lingual lexica[6] by exploiting multilingual Wikipedia to extract the cross-lingual groundings of entities [7]. After obtaining the matching entities, the *top-k graph exploration* is then performed on the graph of the KB for finding the top-k optimal QEGs. The resulting QEGs represent different semantic interpretations of the keyword query. Thus it can help users to refine the query and influence document ranking according to the search intents. More details about our approach to query interpretation can be found in [9].

User Interaction. Different interpretations of the keyword query, i.e., the generated QEGs, are then presented to users for *selecting* the one that fulfills their search intents. The selected QEG can be further *refined*. From an entity in the QEG, users can navigate its description and the connected entities through their relations in the KB, such that they can add additional entities into the QEG or delete unnecessary ones. After that, the entities in the refined QEG constitute the *query entity vector* (*QEV*), where each entry contains the weight of the corresponding entity, which is calculated by the top-k graph exploration algorithm [9] and can also be adjusted by users. These weights will be leveraged for document ranking in the next component.

We consider *user interaction* as beneficial because it enables the interactive query disambiguation and expansion according to users' search intents. Although refinement can be made more precisely on QEGs than on keywords, user interaction is optional in our system. Users can also search the documents directly without interactive query refinement. In this case, the QEG with highest score obtained by the query interpretation component is selected to generate the QEV.

Document Ranking. For document retrieval, the entities in the QEV are used to find relevant documents. However, the documents without the entities in the QEV could also be relevant when they contain entities that are related to the ones in the QEV. Therefore, integrating the related entities into the query can help to cover more complementary information and thus improve the performance of document retrieval. Based on the above observation, we first construct the *expanded query entity vector* (*EQEV*) by *automatically expanding* the QEV with additional related entities.

For each document, we construct the *document entity vector* (*DEV*), where the entries contain the confidence scores of the annotations (i.e., the linked entities of the document), which are generated by our semantic annotation system and stored in the index. It is noted that all the entities in both EQEV and DEV are grounded in the same hub languages such that they serve as the bridge to overcome the language barrier between keyword queries and documents. The semantic similarity between the EQEV and each DEV can be calculated based on standard similarity measures, such as cosine similarity, which is then used for *document ranking*.

[5]The language of the KB to be explored and thus the entities in QEGs grounded in can also be selected by users.

[6]http://km.aifb.kit.edu/sites/xlid-lexica/

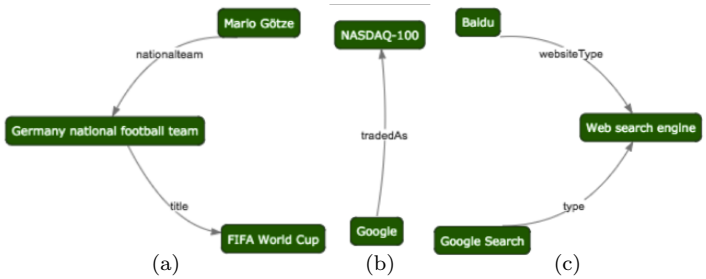

Figure 2: Examples of QEGs for queries (a) *"WM Götze"*, (b) *"online companies of US NDX"* and (c) *"Google 百度"*.

4. DEMONSTRATION

In this section, we demonstrate four major features of XKnowSearch!. A screencast of the demonstration is available at http://km.aifb.kit.edu/sites/XKnowSearch and the online prototype of XKnowSearch! can be accessed at http://km.aifb.kit.edu/services/XKnowSearch.

Query Flexibility. XKnowSearch! supports two search modes: *direct search* and *indirect search*. The direct search mode takes a keyword query and retrieves the relevant documents directly without user involvement in the search process. The indirect search mode provides the opportunity for users to understand the meaning of the query entities and the underlying semantic relations between them yielded by query interpretation, such that users are able to refine and extend the information needs. While the direct search enables users to search in a familiar and convenient manner, the indirect search provides users a *more flexible way* to influence the search process according to their search intents.

Query Disambiguation. Query disambiguation can be performed both automatically and manually. On the one hand, XKnowSearch! *automatically eliminates the ambiguity* of keyword queries by taking advantage of the context, i.e., all candidate query entities, and exploiting the semantic graph of the KB to generate the top-k QEGs. On the other hand, users can also *disambiguate the query manually* by selecting the most appropriate QEG and further refining it. As query interpretations, QEGs are more informative and expressive than keywords such that users can obtain information about not only entities but also relations between them. Consider the keyword query "*WM Götze*", where we assume that the input languages of the keywords are unknown. As shown in Fig. 2 (a), the keyword "*WM*", which could refer to the entity Windows_Mobile in English and FIFA_World_Cup in German[7], has been disambiguated as FIFA_World_Cup based on the relation to Mario_Götze referred to by "*Götze*".

Query Expansion. XKnowSearch! supports keyword search using keywords that match either entities or concepts in their *incomplete* forms, such as aliases, acronyms and misspellings instead of the full names. In addition, the matching concept is automatically expanded into a set of individual entities. Given the keyword query "*online companies of US NDX*" and the top-ranked QEG shown in Fig. 2 (b), it is observed that the alias "*online companies of US*" referring to the concept Internet_companies_of_the_United_States has been resolved to the entity Google, which is listed in NASDAQ-100 referred to by the acronym "*NDX*". Besides

[7]*WM* is the abbreviation of *Weltmeisterschaft* in German, which means *World Cup*.

Figure 3: Examples of the retrieved news articles in different languages for *"Google* 百度*"* and a selected German article.

the role as query interpretation, the resulting QEGs can help users to *manually expand* the query by navigating the KB through entity relations and adding more intended entities that are then used for document retrieval.

Cross-lingual Search. XKnowSearch! enables cross-lingual search in the sense that users can use keyword queries in any language (even in multiple languages) to retrieve multilingual documents, especially in any other languages. The recent progress in cross-lingual technologies is largely due to the increased availability of multilingual data sources on the Web. In this regard, we exploit entities in DBpedia, a multilingual KB, which serve as an interlingua to connect keyword queries and documents across languages. Fig. 2 (c) shows one example of the QEGs generated by our system for the multilingual keyword query "*Google* 百度". Based on that, Fig. 3 illustrates some examples of the retrieved news articles in different languages, where both the query entities (e.g., Google_Inc.) and the additional related entities (e.g., Sergey_Brin), which affect document ranking, are highlighted and linkable to the corresponding resources in the KB.

5. CONCLUSIONS

In this paper, we present XKnowSearch!, a novel entity-based system for multilingual and cross-lingual IR, with the goal of addressing the challenges that traditional keyword search systems mainly suffer from. By leveraging the multilingual KB, namely DBpedia, keyword queries and documents in different languages can be captured on the semantic level to avoid the ambiguity of terms and to bridge the language barriers between queries and documents, where user interaction can also be involved to influence the search process according to the search intents of users.

We believe that this work could complement the term-based document retrieval models and open up new research directions on cross-lingual and entity-based IR. Firstly, it would be interesting to further explore the possibilities of cross-lingual search with the recent initiatives like Wikidata, which tried to make the knowledge bases less language dependent to allow cross-lingual or language independent knowledge access. Secondly, it would be beneficial to query interpretation by also taking into account entity relations expressed in the queries to construct the QEGs. Finally, predicting what type of queries can benefit from the entity-based approach and combining both term-based and entity-based approaches would be a promising direction to pursue.

Acknowledgments.

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 611346.

6. REFERENCES

[1] C. Bizer, T. Heath, and T. Berners-Lee. Linked Data - The Story So Far. *Int. J. Semantic Web Inf. Syst.*, 5(3):1–22, 2009.

[2] M. Färber, L. Zhang, and A. Rettinger. Kuphi - an investigation tool for searching for and via semantic relations. In *ESWC*, pages 349–354, 2014.

[3] J. Hoffart, D. Milchevski, and G. Weikum. STICS: searching with strings, things, and cats. In *SIGIR*, pages 1247–1248, 2014.

[4] X. Liu, P. Yang, and H. Fang. Entexpo: An interactive search system for entity-bearing queries. In *ECIR*, pages 784–788, 2014.

[5] P. N. Mendes, M. Jakob, and C. Bizer. DBpedia: A Multilingual Cross-domain Knowledge Base. In *LREC*, pages 1813–1817, 2012.

[6] J. Pound, P. Mika, and H. Zaragoza. Ad-hoc object retrieval in the web of data. In *WWW*, pages 771–780, 2010.

[7] L. Zhang, M. Färber, and A. Rettinger. xlid-lexica: Cross-lingual linked data lexica. In *LREC*, pages 2101–2105, 2014.

[8] L. Zhang and A. Rettinger. X-lisa: Cross-lingual semantic annotation. *PVLDB*, 7(13):1693–1696, 2014.

[9] L. Zhang, A. Rettinger, and J. Zhang. A knowledge base approach to cross-lingual keyword query interpretation. In *ISWC*, 2016.