A Novel Framework for Related Entities Finding: ICTNET at TREC 2009 Entity Track

Haijun Zhai, Xueqi Cheng, Jiafeng Guo, Hongbo Xu, Yue Liu

Dept of Computer Science and Technology, University of Science & Technology of China, Hefei, China

Institute of Computing Technology, Chinese Academy of Sciences,Beijing, China

{zhaihaijun,guojiafeng,hbxu }@software.ict.ac.cn {cxq,liuyue}@ict.ac.cn

ABSTRACT

This paper addresses the problem of related entity finding, which was proposed in trec 2009. The overall aim of related entity finding (REF) is to perform entity-related search on Web data, which address common information needs that are not that well modeled as ad hoc document search. In this paper, a novel framework was proposed based on a probabilistic model for related entity finding in a Web collection. This model consists of two parts. One is the probability indicating the relation between the source entity and the candidate entities. The other is the probability indicating the relevance between the candidate entities and the topic. Using ClueWeb09 dataset, the experimental evaluations show the effectiveness of our REF framework.

1. INTRODUCTION

With rapid development of World Wide Web, search engines have become an important tool for users to find information they need. Traditional information retrieval systems return a list of documents for users. However, often, users search for specific entities instead of just any type of documents. For instance, when users submit a query "*Michael's teammates while he was racing in Formula 1*", they might expect to find out the names of Michael's teammates. Under related entity finding (REF), users can directly obtain target entities with no need of exploring a large number of documents.

Trec 2009 highlighted the interests in related entity finding. TREC ENTITY task addresses common information needs that are not that well modeled as adhoc document search. The overall aim of this task is to perform entity-related search on Web data, where 31 queries are built (11 queries for training and 20 queries for testing). As an example of the related entity finding task, given the source entity "*Michael Schumacher*", it aims to find all target entities that are related to the source entity "*Michael Schumacher*", where the relation is described by the narrative "*Michael's teammates while he was racing in Formula 1*".

In this paper, we proposed a novel framework for the related entity finding task based on a probabilistic model. Specifically, all candidate entities are ranked by the probability $P(e|Q)$, where e donates candidate entity, Q donates the query which are can be represented by a triple (e^s, R, T), and $P(e|Q)$ donates the probability of Q generating e. $P(e|Q)$ can be computed by multiplying two probability. One is the probability $P(R|e,e^s)$ indicating the relation between the source entity and the candidate entities. The other is the probability $P(e|e^s,T)$ indicating the relevance between the candidate entities and the topic. Note that in triple (e^s, R, T), e^s donates the source entity ("*Michael Schumacher*" in previous example), R donates the relation words ("*teammates*" in previous example), and T donates the type of target entities ("*Person*" in previous example).Using ClueWeb09 dataset, The experimental evaluations show the effectiveness of our REF framework based on the probabilistic model.

2. REF Framework

2.1 Preliminary

Given an input entity, by its name and homepage, the type of the target entity, as well as the nature of their relation, described in free text, find related entities that are of target type, standing in the required relation to the input entity. An example information need, "*find organizations that are sponsoring Kimi Raikkonen*" is formulated as follows:

<*query*>

<*num*>1</*num*>

<*entity_name*>Kimi Raikkonen</*entity_name*>

<*entity_URL*>clueweb09-en0000-00-12345</*entity_URL*>

<*target_entity*>organization</*target_entity*>

<*narrative*>I'd like to know which organizations are Kimi's sponsors.</*narrative*>

</*query*>

2.2 EF Framework

In this section, we describe our REF framework based on a probabilistic model in detail. Our REF framework includes three steps:

(a) **Step 1**: We first build a candidate documents set related to the topic and extract all candidate strings. Specifically, we search the corpus with the narrative of the topic based on BM25 [2] to find out the top 5000 documents as the candidate documents set for the topic. Then, all the anchor texts in the candidate documents set and the titles of the candidate documents are extracted as the candidate strings.

(b) **Step 2**: We build the description document for each candidate string and obtain all candidate entities for the topic. Specifically, for each candidate string obtained in step 1, we extract the top 100 sentences including the candidate string (These sentences are selected from the candidate documents set and the selection method is based on a specific summarization method [4]) from the candidate documents. All the 100 sentences are assembled together as the description document for each candidate string. Note that the candidate strings with no more than five sentences including it are discarded. Finally, applying Stanford NER

Report Documentation Page

Form Approved
OMB No. 0704-0188

Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
NOV 2009		**00-00-2009 to 00-00-2009**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
A Novel Framework for Related Entities Finding: ICTNET at TREC 2009 Entity Track	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Chinese Academy of Sciences,Institute of Computing Technology,Beijing, 100190, China,**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
Proceedings of the Eighteenth Text REtrieval Conference (TREC 2009) held in Gaithersburg, Maryland, November 17-20, 2009. The conference was co-sponsored by the National Institute of Standards and Technology (NIST) the Defense Advanced Research Projects Agency (DARPA) and the Advanced Research and Development Activity (ARDA).

14. ABSTRACT
see report

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**	**Same as Report (SAR)**	**3**	

[1] to the description documents of the candidate strings, candidate entities can be obtained by filtering out the candidate strings with TTS (Target Type Support, TTS is defined by

$$\text{TTS}=\frac{\text{count the tag of the string with target type}}{\text{count the occurrence of the string}})<0.2$$

(c) **Step 3**: We rank all the candidate entities by our probabilistic model and output results. Specifically, we using our probabilistic model to rank all the candidate entities and output their corresponding anchor links (URLs for titles).

In the above, we describe our REF framework based on a probabilistic model. This framework is very effective, the key part of which is the probabilistic model. In the following section, we will introduce the probabilistic model in detail.

2.3 EF Framework

In this section, we describe our probabilistic model. In our REF framework, all candidate entities are ranked by the probability $P(e|Q)$:

$$\begin{aligned} P(e|Q) &\propto P(e,Q) \\ &= P(e,e^s,T,R) \\ &= P(R|e,e^s,T)P(e,e^s,T) \\ &= P(R|e,e^s,T)P(e|e^s,T)P(e^s,T) \\ &\propto P(R|e,e^s,T)P(e|e^s,T) \\ &\approx P(R|e,e^s)P(e|e^s,T) \end{aligned} \quad (1)$$

From Equation (1), we can see that this probabilistic model consists of two parts. One is the probability $P(R|e,e^s)$ which reflects the relation between the source entity e^s and the candidate entity e, which is easy to be computed by counting the co-occurrence of e^s, e and R in the candidate documents set. Note that in Equation (1) we approximate $P(R|e,e^s,T)$ with $P(R|e,e^s)$ under the assumption that the semantics of the target type T can be deduced based on the semantics of the source entity e^s and the relation R.

The other is the probability $P(e|e^s,T)$ which reflects the relevance between the candidate entities and the topic. In the following, we describe our method to estimate the probability $P(e|e^s,T)$. This method includes three steps:

(a) **Step 1**: We build the description document for the narratives of each topic. Specifically, for each narrative, we extract the top 100 sentences from the top 100 documents of the candidate documents set (One sentence is selected from each document and the selection method is based on a specific summarization method [4]). All the 100 sentences are combined together as the description document for each topic.

(b) **Step 2**: We compute the similarity between the description document of each candidate entities and the description document of the topic. Note that cosine similarity is used here.

(c) **Step 3**: We compute the probability $P(R|e,e^s)$ for each candidate entity. Specifically, we using the cosine similarity between the description document of each candidate entity and the description document of the topic normalized with the sum of all the cosine similarity to approximate the probability $P(e|e^s,T)$.

In this method, we use the cosine similarity between the description document of each candidate entity and the description documents of the topic normalized with the sum of all the cosine similarity to approximate the probability $P(e|e^s,T)$, which makes sense since that it has been proved by most research that the snippets are good descriptions for queries [3]. So the cosine similarity between the description document of each candidate entity and the description documents of the topic reflects the relevance between the candidate entity and the topic.

3. Conclusion and Future work

This paper addresses the problem of REF, which aims to perform entity-related search on Web data and address common information needs that are not that well modeled as ad hoc document search. In this paper, a novel framework was proposed based on probabilistic model to entities finding in a Web collection. This model consists of two parts. One is the probability indicating the relation between the source entity and the candidate entities. The other is the probability indicating the relevance between the candidate entities and the topic. Experiments on the ClueWeb09 dataset demonstrated the effectiveness of our method. The average P@10 and NDCG of our method is 0.2350 and 0.2103 respectively. However, much work is stills needed to be conducted. In the future, we will conduct research on the generation of the candidate entities and so on.

4. ACKNOWLEDGMENTS

We thank to Feng Guan and Zeying Peng in the Institute of Computing Technology, Chinese Academy of Sciences. In addition,our work was supported by Key Program of NSFC 60933005，863 Program of China 2007AA01Z441，973 Program of China 2007CB311100.

5. REFERENCES

[1] J. R. Finkel, T. Grenager, and C. Manning. Incorporating nonlocal information into information extraction systems by gibbs sampling. In ACL '05: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, pages 363–370, Morristown, NJ, USA, 2005. Association for Computational Linguistics.

[2] S. R. Nick Craswell, Hugo Zaragoza. Describes application and tuning of okapi bm25f. In In Proceedings of the Fourteenth Text REtrieval Conference (TREC 2005), 2005.

[3] D. Shen, J.-T. Sun, Q. Yang, and Z. Chen. Building bridges for web query classification. In SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval, pages 131–138, New York, NY, USA, 2006. ACM.

[4] C. X. W. G. Zhang, J. and H. Xu. Adasum: an adaptive model for summarization. In In Proceeding of the 17th ACM Conference on information and Knowledge Management, pages 901–910, Napa Valley, California, USA, 2008. ACM.