Wikipedia Article Content Based Query Expansion in IR4QA System

Maofu Liu, Bin Zhou, Liwen Qi and Zilou Zhang
College of Computer Science and Technology, Wuhan University of Science and Technology
liumaofu@wust.edu.cn, zb_zhoubin@163.com

Introduction

◆In information retrieval system, users often submit the query which is a short description by natural language, and they decides the relevance of document not based on semantics of query terms in documents, but existence of query terms.

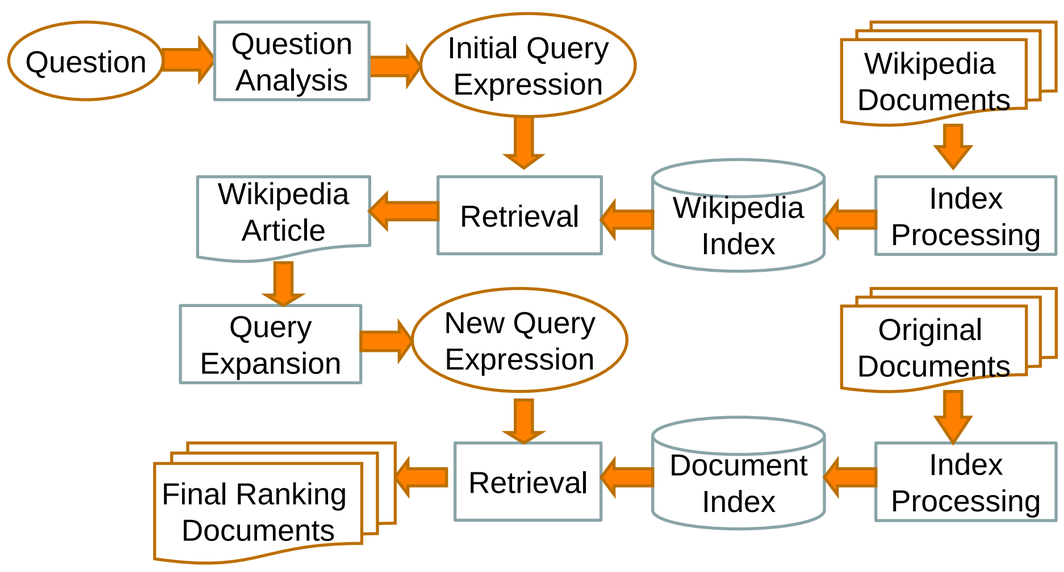
◆The key terms extracted from a question in IR4QA can be different with distinct segmentation strategies.

◆The query expansion is an effective way to solve term mismatch problem by expanding the key terms with a certain number of other related terms in the initial query.

◆Wikipedia is a multilingual, web-based, free-content encyclopedia project, and each of its article provides information to explain the term of the article title.

➢In order to expand the most relevant terms, we make use of related Wikipedia article as the “seed” document of the related question. And then make question expansion based on the most relevant paragraph of the question.

System Architecture



Experiments

RUN	Analysis File	Mean AP	Mean Q	Mean nDCG
WUST-CS-CS-01-T	No	0.2694	0.293	0.4881
WUST-EN-CS-01-T	No	0.1037	0.1206	0.2815
WUST-EN-CS-02-T	WHUQA-EN-CS-03-T.xml	0.1435	0.1564	0.292

Formal run experiment official results (AFTER bug fix)

The system does not achieve a good official result.

◆As our system segment the question into words based on the same dictionary in the index processing module. If the key term does exist in the dictionary, our system may not retrieve the related document.

◆In our “EN-CS” work, we extract English key terms and then translate them into Chinese by Google translation. So the quality of the translation determines the performance of our EN-CS result.

Query Expansion

1. Question Classification

《千里走单骑》和张艺谋是什么关系? → RELATIONSHIP

高仓健是谁? → BIOGRAPHY

2. Article Retrieval

(1) Use different template to extract name entities from different types of question.

(2) Take retrieval in the Wikipedia by the name entities to get the related article. If cannot find, relocate the final article by the most relevant title.

3. Paragraph Location

Question:第76届奥斯卡最佳男主角是谁？

Located Paragraph →

最佳影片
- 魔戒三部曲：王者再临（The Lord Of The Rings: The Return Of The King）

最佳导演
- 彼得·杰克逊(Peter jackson) – 魔戒三部曲：王者再临

最佳男主角
- 辛·潘 (Sean Penn) – 悬河杀机 (Mystic River)

最佳女主角
- 查理丝·花朗 (Charlize Theron) – Monster

最佳男配角
- 添·罗宾斯 (Tim Robbins) – 悬河杀机 (Mystic River)

最佳女配角
- 芮妮·齐薇格 (Renee Zellweger) – 冷山 (Cold Mountain)

4. Query Expansion

Question:李永波和中国羽毛球队是什么关系？
Initial Query Expression:李永波，中国，羽毛球，球队
New Query Expression:李永波，中国，羽毛球，球队，1962年，著名，羽毛球运动，球运，运动员，教练员，辽宁，宁人，现为，国家队，总教练，国家，乒羽，中心，副主任

Term Weight:

$$w(q|Q_{new}) = p \cdot w(q|Q) + k \cdot avg(boost) \cdot \frac{score(q)}{MaxScore} w(q|d)$$

Where $w(q|Q)$ is the weight of key term q in the original query Q, $w(q|d)$ is the weight of q in document d, n is the number of top selected documents, and p and k are experimentally determined positive constants. $boost$ is one factor as a multiplier besides the factors tf and idf to compute the weight of the query key term in the initial query, and the $avg(boost)$ is the average value of them.

Conclusions

◆In order to solve the problems of inappropriate key terms exacted from the initial question and term mismatch, we apply query expansion technique to get more useful key terms for the query based on related Wikipedia article content.

◆we find that when the question types are DEFINITION, BIOGRAPHY, PERSON, ORGANIZATION, LOCATION or DATE, we can find the relevant paragraph easily.

◆The Wikipedia is a document sets of description type, so it performs better for the explanation questions.