Question Analysis and Query Expansion in CS-CS IR4QA

Maofu Liu, Fang Fang, Qing Hu, Jianxun Chen

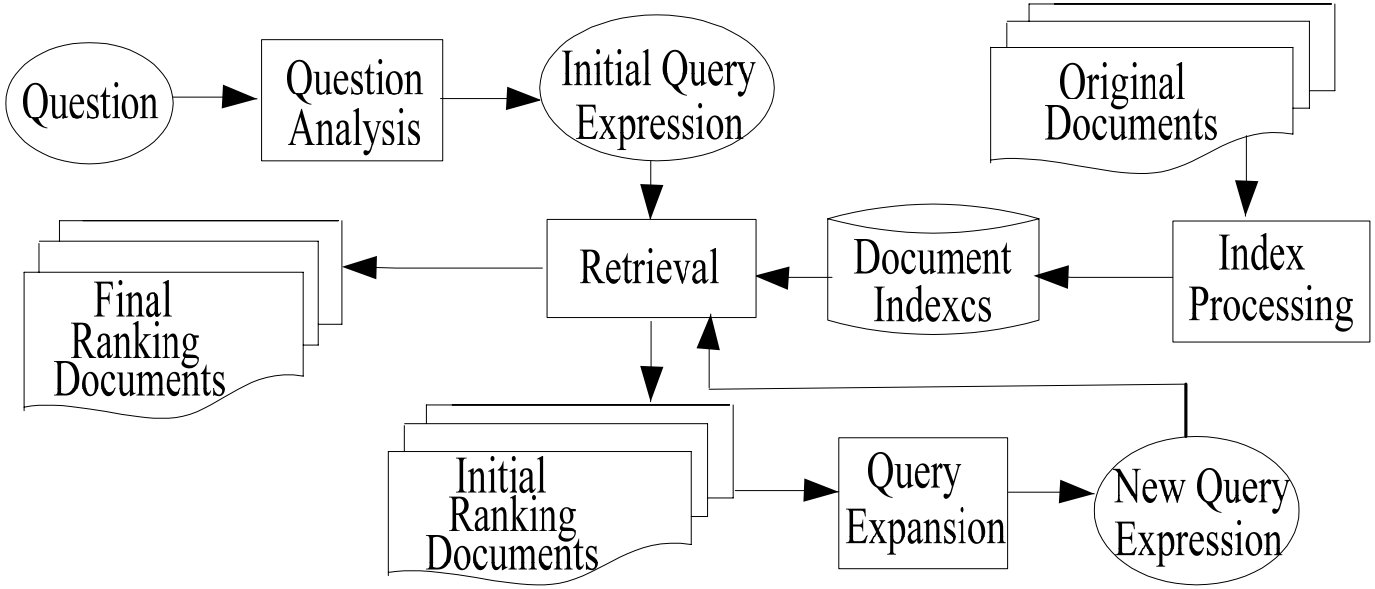
College of Computer Science & Technology

Wuhan University of Science & Technology

Outline

- System Architecture
- Question Analysis
- Query Expansion
- Experiments
- Conclusion

System Architecture



Index processing

- Making segmentation for all of the documents to make word-based index unit and construct inverted index files

Question analysis

- Analyzing the given question to extract key terms from the question to form the query expression for the initial retrieval

Retrieval

- **Retrieval model**
  - **Vector Space Model (VSM)**
- **Two retrieval phases**
  - Initial retrieval
  - Re-retrieval

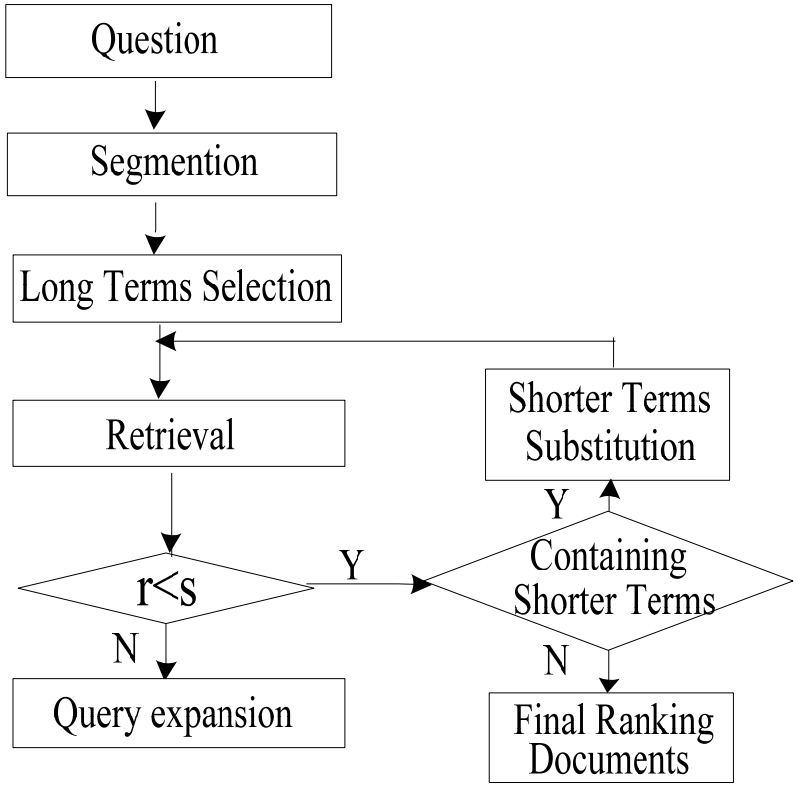
Question expansion

- After the initial retrieval, the PRF approach will be employed to do question expansion

Question analysis

- The purpose of question analysis is to extract key terms from the original question to make up of an initial query expression for the initial retrieval.
- When the user gives a question, the system will segment it into words . But there may be more than one choice for the question segmentation.
- For example, “诺贝尔奖” can be segmented as “诺贝尔奖” only as one term or “诺贝尔+奖” as two terms. The system chose the longer terms in default to do retrieval. If the returned documents number *r* is less than *s*, which denotes the threshold value when use the key terms to search the relevant documents, the key terms can be segmented into the shorter ones and retrieve the original document again.

Question analysis (Cont'd)



Query:列举亚洲金融危机对经济的影响。(List the impact of the Asian financial crisis on the economy.)

Segmentation:
列举,亚洲金融危机(亚洲,金融危机(金融,危机)),对,经济,的,影响.

Long term selection:亚洲金融危机,经济,影响

Query expansion

- 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
- The most well-known query expansion technique is pseudo-relevance feedback (PRF) and it is widely used in the open information retrieval evaluation workshops for improving the retrieval effectiveness
- Two approaches to implement PRF query expansion
  - Co-Occurrence based query expansion approach
  - Metric Correlation based query expansion approach

Co-Occurrence based query expansion approach (CO)

- Deriving from the same idea as LCA(Local Context Analysis)
- Some words appear together in a "window", and the "window" is defined as a document unit in our system.
- We try to seek the co-occurrence of key term *q* in the query and word in the top *n* (*n* is assigned to 50 in our experiment) documents from the initial retrieval. The co-occurrence degree of query term q and word w from each document of the top *n* documents is calculated by the following equations

$$cood(w,q \mid D) = \log(tf(w \mid D) + 1.0) \times \log(tf(q \mid D) + 1.0)$$

CO (Cont'd)

- Employing the following formula to evaluate the co-occurrence degree between the word w and the query Q:

$$cohd(w,q \mid d) = \prod_{q \in Q} (cood(w,q \mid S) + 1.0)$$

- Some words may have high frequency but low distinction, we use the inverse document frequency logarithm to discriminate them and modify the equation as follows

$$f(w,Q \mid C,S) = \sum_{q \in Q} idf(q \mid C) idf(w \mid C) \log(cood(w,q \mid S) + 1.0)$$

- The system uses equation to calculate the co-occurrence degree for each word and chooses the top m words as expanded terms

Metric Correlation based query expansion approach (MC)

- The metric correlation emphasizes the position correlation of words in documents.
- If two words t_u and t_v are in the same document, the distance between two terms is defined as $r(t_u, t_v)$, which denotes the number of words between the two words, then the correlation between q and w can be defined as follows:

$$cood(w, q \mid S) = \frac{\sum_{d \in S} \sum_{w \in d} \sum_{q \in d} \frac{1}{r(q, w)}}{n}$$

- The system compute the score for each word and select the words with higher scores as expanded words

Weight of the expanded terms

- A common way is applying the Rocchio equation directly to compute the weight of each term in the new query
- In our system, we modify this equation, for we consider the score for the expanded term evaluated by the equation also indicates the importance of the terms

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

- Where *w(q|Q)* is the weight of key term *q* in the original query, *w(q|d)* is the weight of *q* in document *D*, *p* and *k* and 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. *score(q)* indicates the score value for each expanded key term and MaxScore is the maximum of them.

Experiments

Evaluation of experiment results

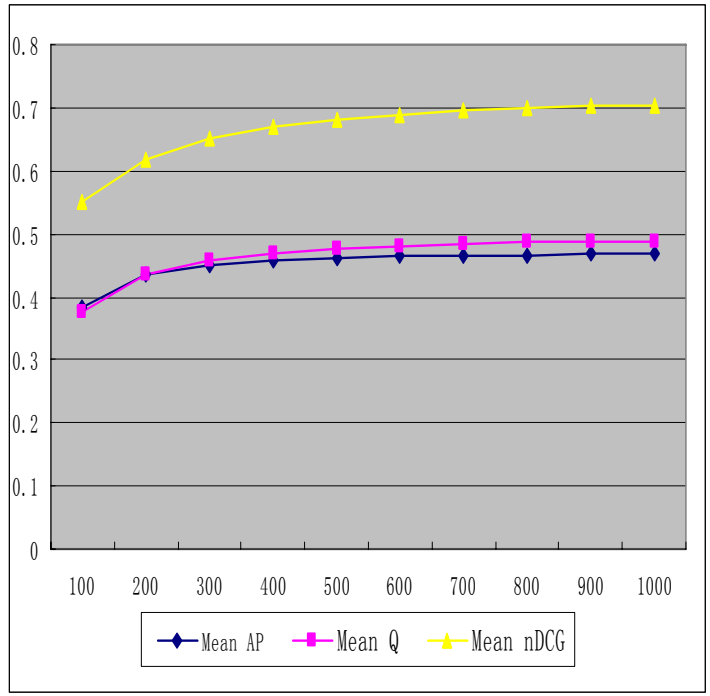
Run	Analysis File	QE	Mean AP	Mean Q	Mean nDCG
01-T	No	MC	0.4316	0.4510	0.6763
02-T	No	CO	0.468	0.4879	0.7051
03-T	CMUJAV-CS-CS-01-T	CO	0.4412	0.4452	0.6031
04-T	Apath-CS-CS-01-T	MC	0.4241	0.4298	0.6139
05-DN	CSWHU-CS-CS-03-DN	CO	0.4720	0.4825	0.6724

"Run" indicates the name of the run file. The suffix "T" indicates the question title and "DN" is the description and the narrative of the question. The "Analysis File" means the question analysis file offered by other participators. "QE" indicates the query expansion approach. In all of the run files, we set 1000 as the limited number of document IDs for each question.

- In order to analyze the relation between the limited number of documents retrieved for each question and the evaluation result of the run file, we get the experimental data as shown in the following table where all of the run files employ our own question analysis means and the CO query expansion approach which the same to the run file 02-T

Evaluation results with different maximal document IDs

MAX	Mean AP	Mean Q	Mean nDCG
100	0.3851	0.3765	0.5507
200	0.4336	0.4337	0.6171
300	0.4503	0.4575	0.6497
400	0.4579	0.4695	0.6699
500	0.4621	0.4767	0.6806
600	0.4647	0.4814	0.6902
700	0.4661	0.4841	0.6964
800	0.4669	0.4858	0.7003
900	0.4676	0.4872	0.7034
1000	0.468	0.4879	0.7051



The evaluation results for ACLIA1-CS-T384 and ACLIA1-CS-T385

Run	Key Terms & boost	AP	Q	nDCG
CSWHU	{(宇宙大爆炸:1.0), (理论:0.3)}	0.7720	0.8342	0.9387
Apath	{(宇宙大爆炸理论:1.0)	0.3987	0.3987	0.5908
CMU	{(宇宙:1.0),(大:1.0),(爆炸:1.0),(理论:1.0)}	0.4044	0.4044	0.5931
NLPAI	{(宇宙大爆炸:1.0), (理论:1.0)}	0.4079	0.4981	0.7847

CSWHU is much better than other approaches for different boost values

Run	Key Terms & boost	AP	Q	nDCG
CSWHU	{(北京大学:1.0), (百年:1.0), (校庆:0.3), (大事:0.3)}	0.0437	0.0475	0.2228
NLPAI	{(北京大学:1.0), (百年:1.0), (校庆:1.0), (大事:1.0)}	0.4283	0.4226	0.6642

CSWHU is much worse than other approaches for different boost values

The key terms are set proper scores for the boost values, it will improve the retrieval performance. On the contray, it will decrease the performance

Conclusion and Future Works

- The system employs the word-unit based index files and the VSM information retrieval model. When a question is submitted, the system analyze the question firstly to select proper key terms and do the first phase retrieval. Then it apply the PRF approach to do question expansion to get more relevant key terms and do the second phase retrieval.
- By experiments, we find that these techniques can reduce the impact of the term mismatch and enhance the retrieval performance. But there are still some important factors which are set inappropriately in the system, such as the better index unit for the index files, the ignorant of part-of-speech feature and so on. Our future work will make researches on these issues.

Acknowledgment

- The work in this paper was supported partially by a Special Research Grant from Wuhan University of Science and Technology (No. 50003301) and partially by Research Grant from Hubei Provincial Department of Education (No. 500064).

- For further information, please contact liumaofu@wust.edu.cn or fangfang0402@126.com