AUTOMATICALLY EXPANSION OF VIETNAMESE QUERY ON THE INTERNET: STUDYING AN APPROACH AND IMPLEMENTING A DEMO

ABSTRACT: Information retrieval (IR) is a very interesting field of study. Although there are many researches in searching technique, but there are few study relating to improve the effectiveness of retrieval, especially in Vietnamese. Nowadays, Vietnamese text retrieval problem is just at the beginning phase, so making more researches in order to create new effective search engine as well as apply achievements of IR in other languages into Vietnamese is very important. This paper explores some methods, especially automatic query expansion method, to improve the effectiveness of a search engine and some problems needed to resolve when building Vietnamese search engine. From those explorations, planning a solution for making Vietnamese search engine effectively.

1. General

The purpose of an information retrieval is to search for all related documents in a collection according to a certain user query. Simply, we can consider of this system as a black-box as follow :

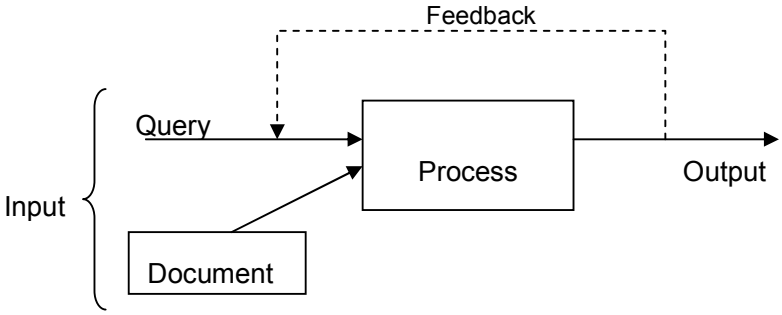

Diagram 1. Logic process of an information retrieval system

2. Some approaches in improving searching effectiveness

To improve the effectiveness of an information system, researchers have some approaches as follow :

+ Expanding the data store/ warehouse using for searching.
+ Improving data structure and searching algorithm
+ Improving speed by using modern techniques of hardware & software
+ Improving inter-relation between system and user (appearance/ feedback)
+ Expanding user query in order to get more suitable results.

In the last approach, many solutions have been researched and applied for more than 3 decades. They can be summarized as follow :

Solution	Source of vocabulary	Technique	Advantage	Disadvantage
1. Base on syntax	Vocabulary/ Grammar Dictionary	Comparison/ Matching	Simple/ base on built-in dictionary	Effectiveness is not good.
2. Base on relation	Relation tree (synonym/ antonym dictionary)	Keyword matching	Get more related results.	Not enough information about global relation.
3. Automatically word classification	Word classes for certain subject.	Statistic, ranking	Simple	Depend on word classes, not good
4. Automatically document classification	Keywords in a class of document for a certain subject.	Statistic, ranking	Better results.	Expensive, depend on doc SimilarityThreshold
5. Base on single word	Synonym dictionary (word-word)	Natural Lang. Processing	Improving effectiveness.	Not pay attention to word order
6. Base on phrase	CombinedDictionary (word-phrase)	Natural Lang. Processing	Improving most effectiveness	Not easy to Identify phrases
7. Base on feedback	Previous results	Natural Lang. Processing	Simple, effective	Hard to identify expand parameter

Table 1. Some solutions for expanding user query.

3. Automatically expansion of query

Automatically expansion of quey is one of some methods in improving the effectiveness of information retrieaval. This method uses information about context, syntax, vocabulary… to find out some more words suitable for expanding. These words could be taken from some sources, such as : synonym dictionary, feedback documents, previous information, recent results,… Then, these words could be added in the original query and do the search again with new query.

The results will be returned in form of list of ranked documents, based on its similarity to the query. The similarity rate between a document d_j and a query q_k is as follow :

$$sim(d_j, q_k) = \frac{d_j^T q_k}{\| d_j \| . \| q_k \|} \qquad (3.4)$$

with ||.|| is Euclide standard of a vector.

Although, there are many ways in expanding a query, they all have same questions to answer :

(1) Which words (phrases) should be added to expanding query ?

(2) How to choose these words and how is the results when we change some of technique parameters.

4. Expanding Vietnamese query : some problems

There are some problems should be solved when applying above method in expanding Vietnamese query :

- Vietnamese words spitting (most compound words)
- Vietnamese literature and font standard (Unicode/ Other codes).
- Identifying Vietnamese word classification (unclearly).
- Open source code Vietnamese dictionary (not yet)
- Vietnamese query with/ without accent signs.

Some of these problems have been solved somewhere, but not enough, and should be considered completely in some more researches.

Applying expanding method to Vietnamese query, with certain attention to above problems, we can follow these suggested steps in following diagram :

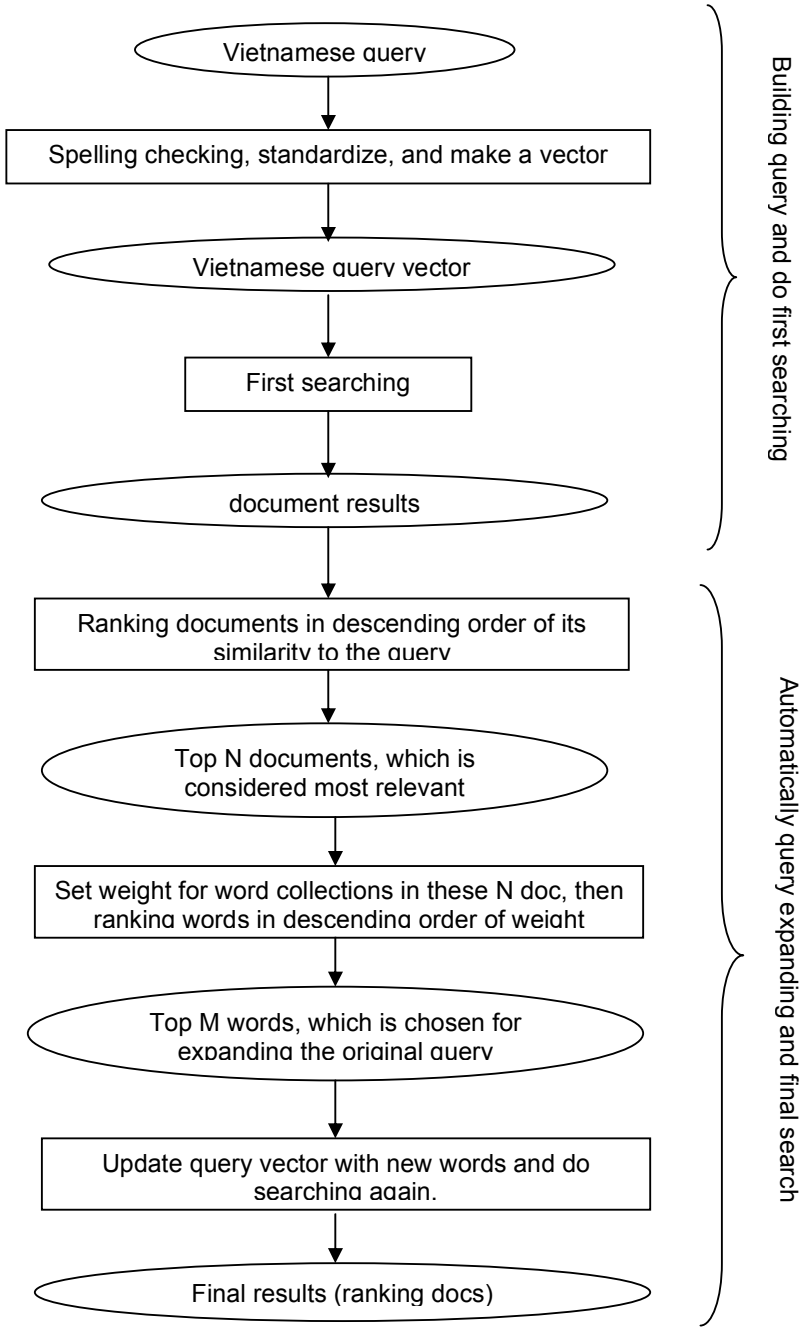

Diagram 2. Suggested steps for expanding Vietnamese query

5. Experimentation and remark

5.1. Experiment Description :

- Experiment 1 : It's used to test system effectiveness with some kind of query: 1, 2, 3 word-length query. Do searching with 10 random 1 word-length queries, save the results and remarks. Do again with 2, 3 word-length queries. Compare the results.
- Experiment 2 : It's used to compare demo system and 3 other search engines : google.com.vn; panvietnam.com; vinaseek.com. Do searching with 10 random queries on above systems. Save the results and compare.
- Experiment 3 : It's used to compare demo system effectiveness before and after expanding. Do searching with 10 random queries on demo system, save added words. Do searching again with google.com.vn with original query and expanding query. Save the results and compare.

5.2. Remarks :

- These 3 experiments show that demo system of expanding Vietnamese query is fairly effective. Although, the average precision is not as high as expected. Some reasons may be the database using for testing is not large enough, some steps depend on previous works, lack of time...
- In comparison to google.com.vn, the testing system is far below, but in comparison to other Vietnamese search engine, such as panvietnam.com, the testing system is better.
- When using expanding queries on google.com.vn, the results is more relevant, and precision rate is higher than before.

6. Conclusion

6.1. Some results :

- Studying approaches :

  This thesis introduces some problems of an information retrieval system, and explores some methods for improving the effectiveness of information retrieval system. One of some approaches is chosen to study is query expanding automatically.

  Moreover, the thesis also apply some small modifications for expanding Vietnamese queries based on feedback information.

- Implementing a demo :

  With these explorations, a suggestion for implementing a Vietnamese search engine using this approach. It is the first ideas for building a real system in the future.

  Besides, some experiments are conducted to test the effectiveness of demo system and to compare with other system. In general, the results support the suggested approach.

6.2. Future Suggestions

Some suggestion for further researches is as follow :

- Consider weighting and ranking system on the improvement.
- Using feedback information from user, context, previous search
- Improving Vietnamese language processing…
- Applying multi-language searching techniques.
- Using parallel processing, hybrid computing…
- Combine with data-mining and knowledge exploring…

REFERENCES

[1] Diệp Quang Ban, Hoàng Văn Thung (2001), ***Ngữ pháp tiếng Việt***, NXB. Giáo Dục, Hà Nội.

[2] Đinh Điền, Hoàng Kiếm, Nguyễn Văn Toàn (2001), ***Vietnamese Word Segmentation***, National University of HCMC, Vietnam.

[3] Claudio Carpineto, Renato De Mori, Giovanni Romano, Brigitte Bigi (2001), ***An Informatic-Theoretic Approach to Automatic Query Expansion*** – ACM Press, USA.

[4] Stefan Klink, Armin Hust, Markus Junker, và Andreas Dengel (2002), ***Improving Document Retrieval by Automatic Query Expansion Using Collaborative Learning of Term-Based Concepts*** – German Research Center for AI., Germany.

[5] C. J. Van Rijsbergen (2000), ***Information Retrieval***, Department of Computing Science, University of Glasgow, UK.

[6] Some Internet websites : http://www.dactrung.net/ ; http://www.google.com.vn ; http://trec.nist.gov/ ; http://www.panvietnam.com/ ; http://www.vinaseek.com/ ;