ISSN: 2278 – 7798
International Journal of Science, Engineering and Technology Research (IJSETR)
Volume 2, Issue 4, April 2013

Classification and Comparative Analysis of Passage Retrieval Methods

Dr. K. Saruladha, C. Siva Sankar, G. Chezhian, L. Lebon Iniyavan, N. Kadiresan
Computer Science Department, Pondicherry Engineering College, Puducherry, India
siva_baba@pec.edu

***Abstract*— In Information Retrieval and Natural Language Processing (NLP), Question Answering (QA) is the task of automatically answering a question posed by humans in natural language. QA is a specialized form of Information Retrieval (IR). Passage retrieval, which aims for extracting the relevant passage inside a text by computational means, has long been studied in IR and has been an important component in QA Systems. The passage retrieval module acts as an intermediate stage between document retrieval module and answer extraction module. This paper provides a brief discussion on the various passage retrieval systems and the methods that has been used in these systems. In addition the paper also describes the limitations of the existing passage retrieval methods.**

***Index Terms*— Information Retrieval (IR), Natural Language Processing (NLP), Question Answering (QA) system, Passage Retrieval (PR).**

I. INTRODUCTION

Information Retrieval (IR) is the process of locating and retrieving documents relevant to a user's information need from a collection of documents. The user's information need is presented to the IR system as a query which usually consists of a string of words. The IR system uses a matching mechanism to decide how closely a document is related to the query. Passage retrieval has long been studied in Information retrieval. The aim is to perform fine grained, targeted information retrieval in response to user query posed in natural language rather than retrieving a list of potentially relevant documents. Passage retrieval is an essential component in QA systems. Most current QA systems employ a pipeline structure that consists of several modules to get short and precise answers to the users questions.

A typical pipeline of a Question Answering System consists of three distinct phases:

- Question classification
- Information retrieval or document processing
- Answer extraction.

Question classification is the first phase which classifies user question, derives expected answer types, extract keywords, and reformulates a question into semantically equivalent multiple questions.

Document processing uses search engines to identify the documents or paragraphs in the document set that are likely to contain the answer.

Answer extraction is a final component in question answering system, which searches the passage for correct answers.

Passage retrieval module acts as an intermediate stage between the document retrieval phase and answer extraction phase. Passage retrieval greatly affects the performance of the QA system. If the passage retrieval module retrieves too many irrelevant passages, the answer extraction is likely to fail to extract the exact answer due to too much noise. Given a document, the passage retrieval system extracts and displays the text which best match the corresponding query. Thus regardless of the length of the document, the user can tell at a glance whether a given document is likely to answer their information need. Another usage of passage retrieval is as answer reporter or answer indicator. If the passages are retrieved and displayed, regardless of the document they are in, according to their sole resemblance to, say, a question, they are likely to produce user with direct answer (*answer reporter*) or with an indication that the answer can be found in a particular document (*answer indicator*). Thus the passage retrieval can thus provide facts to users on a particular subject.

The rest of the paper is organized as follows in Section II discusses the architecture of the passage retrieval system, Section III discuss the classification of passage retrieval system and a brief explanation of the methods used in it. Section IV discusses on the limitations of the passage retrieval systems.

Manuscript received March 22, 2013.
***Dr. K. Saruladha**, Computer Science Department, Pondicherry Engineering College, (e-mail: charuladha@pec.edu). Puducherry, India, 9442396080. **C. Siva Sankar**, Computer Science Department, Pondicherry Engineering College, (e-mail: siva_baba@pec.edu). Puducherry, India, 9003564313. **G.Chezhian**, Computer Science Department, Pondicherry Engineering College, (e-mail: chezhian_g@pec.edu). Puducherry, India, 9789701710. **L. Lebon Iniyavan**, Computer Science Department, Pondicherry Engineering College, (e-mail: smart_lebon@pec.edu). Puducherry, India, 9600831726. **N. Kadiresan**, Computer Science Department, Pondicherry Engineering College, (e-mail: nkadiresan@pec.edu). Puducherry, India, 9944524613.*

II. PASSAGE RETRIEVAL SYSTEM ARCHITECTURE

The architecture of Passage Retrieval (PR) system [2] consists of several components as shown (Figure 1).

Components of the Passage Retrieval Systems

The major task of the passage Retrieval is to locate passages in the corpus of documents. The following are the sequence of steps involved in passage retrieval is:

- **Query**
  The user will post a query or question to get the relevant answer or passage.

- **Query Analyzer**
  The query analyzer phrases the question into subject, verb, object etc. It analyzes the question posted by a user to form a query for document retrieval. It also used to improve the performance of the QA system.

- **Query Expansion**
  Query expansion is the process of reformulating a seed query to improve retrieval performance in information retrieval operations. In the context of web search engines, query expansion involves evaluating a user's input and expanding the search query to match additional documents.

- **Document Retriever**
  Document Retriever retrieves potentially relevant documents from the document collection (corpus) for a given query. This process reduces the corpus to a manageable set of documents for additional processing.

- **Passage Retriever**
  Passage Retriever retrieves relevant passages from listed documents for a given query. This component further reduces document set to a small set of passages, which are later processed to extract answers.

- **Answer Extractor**
  Answer Extractor looks in retrieved passages for answer to user's natural language question. Finally it retrieves the exact answer for a query or question.

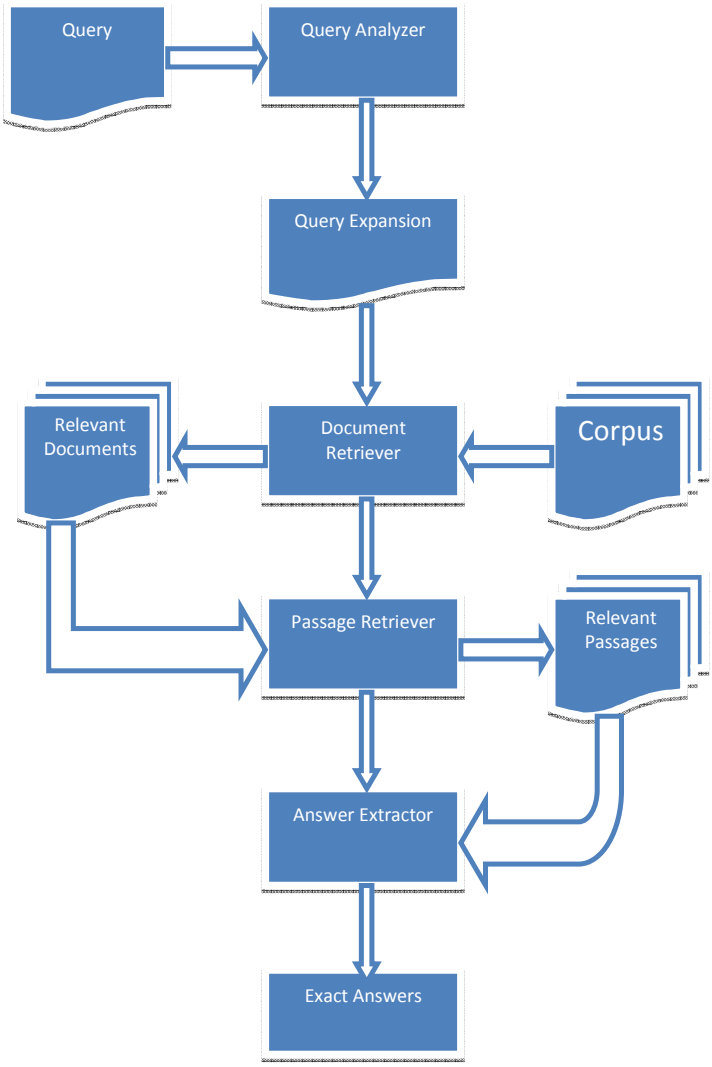

Figure 1: Passage Retrieval System Architecture

The next section classifies the existing Passage Retrieval methods into three major categories

- Semantics Based Passage Retrieval Method
- Statistical Based Passage Retrieval Method
- Structural Based Passage Retrieval Method

Each of these passage retrieval systems is explained in detail in following sections.

III. CLASSIFICATION OF PASSAGE RETRIEVAL METHODS

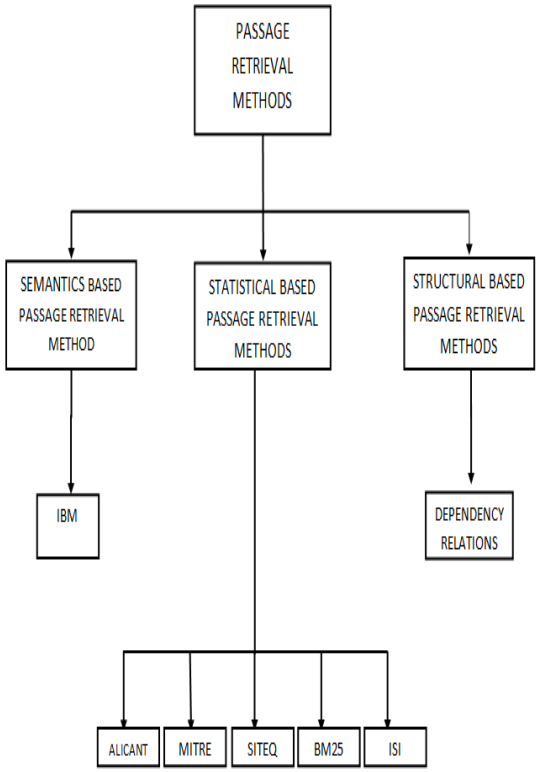

Figure 2: Classification of Passage Retrieval Methods

3.1 SEMANTIC BASED PASSAGE RETRIEVAL METHODS

Semantic based approach leads to the retrieval of relevant passages and accuracy in answering. The semantic based approach uses linguistic dictionaries for retrieving passages.

3.1.1 IBM [6, 5]

- Computes the series of distance measures for the passage
- The "matching words measure" sums the idf values of words that appear in both the query and the passage
- The "thesaurus match measure" sums the idf values of words in the query whose WordNet synonyms appear in the passage.
- The "mismatch words measure" sums the idf values of words that appear in the query and not in the passage.
- The "dispersion match measure" counts the number of words in the passage between matching query terms

3.2 STATISTICAL BASED PASSAGE RETRIEVAL METHODS

In the Statistical method, the passages are retrieved based on the term density or tf-idf values or cosine similarity that appears most common with the query and the passages. In this method, each paragraph in a document as being in one of the two states: "relevant" and "irrelevant".

3.2.1 ALINCATE [2, 3]

- Computes the non-length normalized cosine similarity between query terms and the passage.
- Score is computed taking into account the number of appearances of a term in the passage and in the query along with their idf values.

3.2.2 MITRE [11]

- The word overlap algorithm presented by Light *et* al.[13] counts the number of terms a passage has in common with the query.
- Each sentence is treated as a separate passage.

3.2.3 SITEQ [4]

- Computes the score of an n-sentence passage by summing the weights of the individual sentences.
- Sentences are weighted based on query term density.

3.2.4 BM25 [5, 6]

- Counts the number of terms a passage has in common with the query.
- Score is computed based on the query terms and the document length.

3.2.5 ISI [1]

- Rank sentences based on the similarity to the question.
- Weighing features: Exact match of proper names, match of query terms and match of stemmed words.

3.3 STRUCTURAL BASED PASSAGE RETRIEVAL METHODS

The Passage Retrieval algorithms discussed in the previous section do not take structure of the sentences or passage into consideration. The following methods consider the structure of the sentence into consideration.

3.3.1 DEPENDENCY RELATION METHOD [12]

- A statistical technique for measuring the degree of match of pertinent relations in the candidate sentences with their corresponding relations in the question.
- Sentences that have similar relations between question terms are preferred.
- All single relations between any two terms (or nodes) in the parse tree is treated as a *relation path*.

IV. LIMITATIONS OF THE PASSAGE RETRIEVAL METHODS

The table below shows the methods and the limitations of these existing passage retrieval algorithms:

S. No.	PASSAGE RETRIEVAL ALGORTIHMS	DESCRIPTION	LIMITATIONS
1.	MITRE[11]	• Counts the number of terms a passage has in common with the query. • Each sentence is treated as a separate passage.	• Simple • Ignores semantics • Dependency between words is ignored • Context not considered
2.	BM25[9,10]	• Counts the number of terms a passage has in common with the query. • Score is computed based on the query terms and the document length	• Simple • Ignores semantics • Dependency between words is ignored • Context not considered
3.	Dependency Relation [7,8]	• A statistical technique for measuring the degree of match of pertinent relations in the candidate sentences with their corresponding relations in the question. • Sentences that have similar relations between question terms are preferred. • All single relations between any two terms (or nodes) in the parse tree is treated as a *relation path*	• Ignores semantics
4.	IBM [5,6]	• Computes the series of distance measures for the passage	• Dependency between words is ignored • Context not considered

		• The "matching words measure" sums the idf values of words that appear in both the query and the passage • The "thesaurus match measure" sums the idf values of words in the query whose WordNet synonyms appear in the passage. • The "mismatch words measure" sums the idf values of words that appear in the query and not in the passage. • The "dispersion match measure" counts the number of words in the passage between matching query terms	
5.	SiteQ[4]	• Computes the score of an n-sentence passage by summing the weights of the individual sentences. • Sentences are weighted based on query term density.	• Simple • Ignores semantics • Dependency between words is ignored • Context not considered
6.	Alicante[2,3]	• Computes the non-length normalized cosine similarity between query terms and the passage. • Score is computed taking into account the number of appearances of a term in the passage and in the query along with their idf	• Simple • Ignores semantics • Dependency between words is ignored • Context not considered

		values.	
7.	ISI[1]	• Rank sentences based on the similarity to the question. • Weighing features: Exact match of proper names, match of query terms and match of stemmed words.	• Simple • Ignores semantics • Dependency between words is ignored • Context not considered

V. COMPARISON OF DATA SETS, DATA SOURCES, TOOLS USED AND EXPERIMENTS IN VARIOUS PASSAGE RETRIEVAL ALGORITHMS

Passage Retrieval Algorithms	**TREC Data sets**	**Data Sets**	**Data Sources**	**Tools Used**	**Experiments**
MITRE	TREC 8 TREC 9	1. The TREC 8 Question Answering Track Data Set 2. The CBC Reading Comprehension Data Set	500,000 Newswire Documents 259 Documents or Stories	-	-

Bm25	Okapi at TREC 3 Okapi at TREC 4	1. The TREC 3 Question Answering Track Data Set 2. The TREC 4 Question Answering Track Data Set	Collection of Documents Collection of Documents	Okapi Software	City's TREC Experiments and Interactive Track Experiment
IBM	TREC 9 TREC 10	1. The TREC 9 Question Answering Track Data Set 2. The TREC 10 Question Answering Track Data Set	Collection of Documents Collection of Documents	-	-
SiteQ	TREC 10	1. The TREC 10 Question Answering Track Data Set 2. Web Track	The document collection consists of the following six data sets: AP newswire, Wall Street Journal, San Jose Mercury News, Financial Times, Los Angeles Times, and Foreign Broadcast Information Service.	-	-

Alicante	TREC 10	The TREC 10 Question Answering Track Data Set	Collection of Documents	SUPAR NLP tool	-
ISI	TREC 10	The TREC 10 Question Answering Track Data Set	Collection of Documents	-	-

Table 2: Comparison of Data Sets, Data Sources, Tools used and Experiments in various passage retrieval algorithms

VI. CONCLUSION

In this paper we made comparative analysis on different passage retrieval methods and its limitation. Although passage retrieval is an important component of question answering systems, end-to-end performance depends on a variety of other factors. Measuring passage retrieval performance semantics match of query terms are important in passage ranking. In addition, our investigations of the effects of document retrieval systems suggest that the interaction between document retrieval and passage retrieval is just as important.

REFERENCES

[1] E. Hovy, U. Hermjakob and C.-Y. Lin, "The use of external knowledge in factoid QA", In Proceedings of the Tenth Text REtrieval Conference (TREC 2001), 2001.

[2] F. Llopis and J. L.Vicedo, "IR-n: A passage retrieval system at CLEF-2001", In Proceedings of the Second Workshop of the Cross-Language Evaluation Forum (CLEF 2001), 2001.

[3] J. L. Vicedo and A. Ferr´ andez, "University of Alicante at TREC-10", In Proceedings of the Tenth Text REtrieval Conference (TREC 2001), 2001.

[4] G.G. Lee, J. Seo, S. Lee, H. Jung, B.H. Cho, C. Lee, B.K. Kwak, J. Cha, D. Kim, J. An, H. Kim and K. Kim, "SiteQ: Engineering high performance QA system using lexico-semantic pattern matching and shallow NLP," In Proceedings of the Tenth Text Retrieval Conference, pp. 442-451, January 2001.

[5] A. Ittycheriah, M. Franz and S. Roukos, "IBM's Statistical Question Answering System – TREC-10," In Proceedings of the Tenth Text Retrieval Conference, 2001.

[6] A. Ittycheriah, M. Franz, W.J. Zhu and A. Ratnaparkhi, "IBM's statistical question answering system," In Proceedings of the 9th Text REtrieval Conference (TREC-9), 2001.

[7] C.Clarke, G.Cormack, D.Kisman and T.Lynam, "Question Answering by passage selection(Multitext experiments for TREC-9)," In Proceedings of the Nineth Text Retrieval Conference(TREC-9), 2000.

[8] C.Clarke, G.Cormack and E.Tudhope, "Relevance ranking for one to three term queries," Information Processing and Management, vol.36, pp. 291-311, 2000.

[9] S.E. Robertson, S.Walker, M.Hancock-Beaulieu, M.Gatford and A.Payne, "Okapi at TREC-4," In Proceedings of the 4th Text Retrieval Conference(TREC-4), 1995.

[10] S.E. Robertson, S.Walker, S.Jones, M.Hancock-Beaulieu and M.Gatford, "Okapi at TREC-3," In Proceedings of the 3th Text Retrieval Conference(TREC-3), 1994.

[11] M.Light, G.S.Mann, E.Riloff and E.Breck, Analyses for elucidating current question answering technology, "Special Issue on Question Answering," Journal of Natural Language Engineering, pp.325-342, Fall-Winter 2001.

[12] Hang Cui, Renxu Sun, Keya Li, Min-Yen Kan and Tat-Seng Chua, "Question Answering Passage Retrieval Using Dependency Relations," Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 400-407, 2005.

[13] Stefanie Tellex, Boris Katz, Jimmy Lin, Aaron Fernandes and Gregory Marton, "Quantitative Evaluation of Passage Retrieval Algorithms for Question Answering," Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 41-47, July 2003.

[14] Wei Xu, Ralph Grishman, Le Zhao, "Passage Retrieval for Information Extraction using Distant Supervision" Proceedings of the 5th International Joint Conference on Natural Language Processing, pages 1046–1054,Chiang Mai, Thailand, November 8 – 13, 2011.

[15] F. Song and B. Croft, "A general language model for information retrieval," In Proceedings of the 1999 ACM SIGIR Conference On Research and Development in Information Retrieval, pp. 279-280, 1999.

[16] Jianfeng Gao, Jian-Yun Nie, Guangyuan Wu, Guihong Cao, "Dependency Language Model for Information Retrieval," Proceedings of the 27th Annual International ACM SIGIR Conference On Research and Development in Information Retrieval, pp. 170-177, 2004.