A Context-Aware Framework for Semantic Indexing of Research Papers

Maryam Tayefeh Mahmoudi[1,2], Fattaneh Taghiyareh[1], Koushyar Rajavi[1], Mohammad Saleh Pirouzi[1]

[1)] School of ECE, College of Engineering
University of Tehran, Iran
[2)] Knowledge Management & E-Organizations Group
IT Research Faculty, Research Institute for ICT (ITRC), Tehran, Iran
Emails: {mahmodi@itrc.ac.ir, ftaghiyar@ut.ac.ir, k.rajavi@ece.ut.ac.ir, s.pirouzi@ut.ac.ir}

***Abstract*—Automatic indexing and annotation of publications have a significant role in retrieving and processing required papers from the massive amount of existing papers in the databases. In this paper, a framework for indexing research papers based on domain ontology is represented. The domain ontology, which is constructed for this purpose, is on agent science and technology. The initial step to indexing is to recognize the major concerns and the basic constituents in the title of papers, which has been accomplished through proposing a few NLP-based rules. To annotate each paper, the mentioned ontology and WordNet are employed. Experimental results on about 155 research papers lead us to estimate that our framework is capable of semantic indexing in about 80 percent of the situations. Since we have considered the ontology separately from the constituents of the whole system, the proposed framework is domain-independent and can be applied to any other domain ontology.**

Keywords-Indexing; domain ontology; research paper; WordNet; incremental learning.

I. INTRODUCTION

Semantic Indexing has an influential role in managing tremendous amount of publications in databases. The goal of semantic indexing is to offer more effective search and categorization services. There exist various methods for this purpose such as Latent semantic indexing (LSI) and concept indexing (CI), which are among information retrieval techniques [1, 2]. Although they have empirical success, they suffer from the lack of interpretation for the low-rank approximation and, consequently, the lack of controls for accomplishing specific tasks in information retrieval [3]. To overcome the existing deficiencies, there is a tendency to more potential schemes like ontology and semantic web. Domain ontology seems to be an appropriate tool for supporting indexing methods which can be widely used for knowledge and content processing applications [4, 5, 6]. In the meantime, the construction of domain ontology relies on domain modelers and knowledge engineers that are typically overwhelmed by the potential size, complexity and dynamicity of a specific domain [7]. To overcome the barrier of constructing exhaustive domain ontology, annotating or indexing may again be an appropriate alternative to enable the ontology with the potential of learning [8]. Thus, close examination of the issue reveals that indexing plays a major role both in publications' storage management and consequently incremental ontology learning.

Taking the rapid growth in the number of research papers into account, turns into deployment of appropriate indexing method for facilitating both storage and retrieval purposes [9].

In this paper, we propose a context-aware framework for semantically indexing research papers based on domain ontology and NLP-based rules. The domain ontology which is constructed for this purpose is on agent science and technology issue, while the NLP-based rules are achieved through processing huge amount of research papers' titles. The main reason that titles of publications are considered as the basis of indexing is that they are informative enough that there is no need to process the whole text. Determining major concern and basic constituent behind the title leads into semantically indexing each paper. By major concern, we mean the major objective and concern of the title that illustrates why and for what reason it is under consideration, while by basic constituent we mean how to realize the major concern, by applying special tools or means [10, 11]. Having a review on existing approaches using major concern and basic constituent reveals the potential of these concepts in indicating the main objective behind the whole text [12, 13]. To automate the above mentioned process, some NLP-based rules are proposed. It is no doubt that by matching the major concern and basic constituent with the existing concepts in the domain ontology, annotation will be accomplished. It is to be added that WordNet can also be a supportive tool for finding the closest concepts in an unmatched cases.

The rest of the paper is organized as follows: Section 2 reviews some of the previous works which have been done in the area of indexing and annotating. Section 3 describes our suggested framework. In Section 4, experimental results are analyzed and Section 5 sketches out the conclusion and future works.

 ISBN: 978-1-61208-181-6

II. RELATED WORK

Due to the rapid growth in number of publications, organizing papers and documents in a certain database has become more important than before, especially for store, search and retrieval purposes [14].

In the meantime, there exist various indexing or annotating methods which are discussed as follows:

Some focus on phrase-based document similarity via index graph model. This method has the potential of detecting any-length phrase match from the current document to all the previously seen documents in the data set by just scanning it and extracting the matching phrases from the document index graph [15]. Index-Filter is another method, which uses indexes built over the document tags to avoid processing large portions of the input document [16]. In addition to data structure for indexing XML documents based on relative region coordinates which describe the location of content data in XML documents is also mentionable [17]. With respect to managing the large spatial ontologies, spatial index for improving the efficiency of the spatial queries are deployed [18].

In addition to the methods discussed above, Latent Semantic Indexing (LSI) or Latent Semantic Analysis (LSA) and also concept indexing (CI) are mentionable. Those methods improve the detection of relevant documents on the basis of terms found in queries [19]. The most challenges to LSI focused on scalability and performance. LSI requires relatively high computational performance and memory in comparison to other information retrieval techniques [20].

PubSearch uses a citation based retrieval system [14] which generates a web citation database from online scientific publications that are available over the internet. Random indexing is another method which is based on an incremental word space model [21]. The basic idea of Random Indexing is to accumulate context vectors based on the occurrence of words in documents.

It is not to be disregarded that ontologies have significant role in semantic annotation, too [5, 22]. They are being widely used in information retrieval (IR) either for performing semantic indexing of documents or to produce a better organization of retrieved documents [23]. In this respect, document indexation methods, specifically in large-scale web search engines, support the retrieval of documents that might contain some parts related to the query [24]. Linguistic annotation is also an important field in natural language processing that involves classification of text into a predefined set of values [25]. Improving the semantic capability of ontology-based indexing method by major concern and basic constituent is our concern in this paper.

III. SUGGESTED FRAMEWORK

A. The overall Structure of Proposed Framework

As it has been mentioned before, in large-scale databases of research papers, applying a well–defined indexing method plays a significant role to retrieve desired papers. In this respect, we propose a context-aware framework that seems to be capable enough to facilitate such a process. Figure 1 illustrates the details of proposed framework.

As it is illustrated in Figure 1, each time that a paper is uploaded to the database, it is necessary to be indexed. For this purpose, extracting the title of the paper and parsing it is required in order to find major concern and basic constituent.

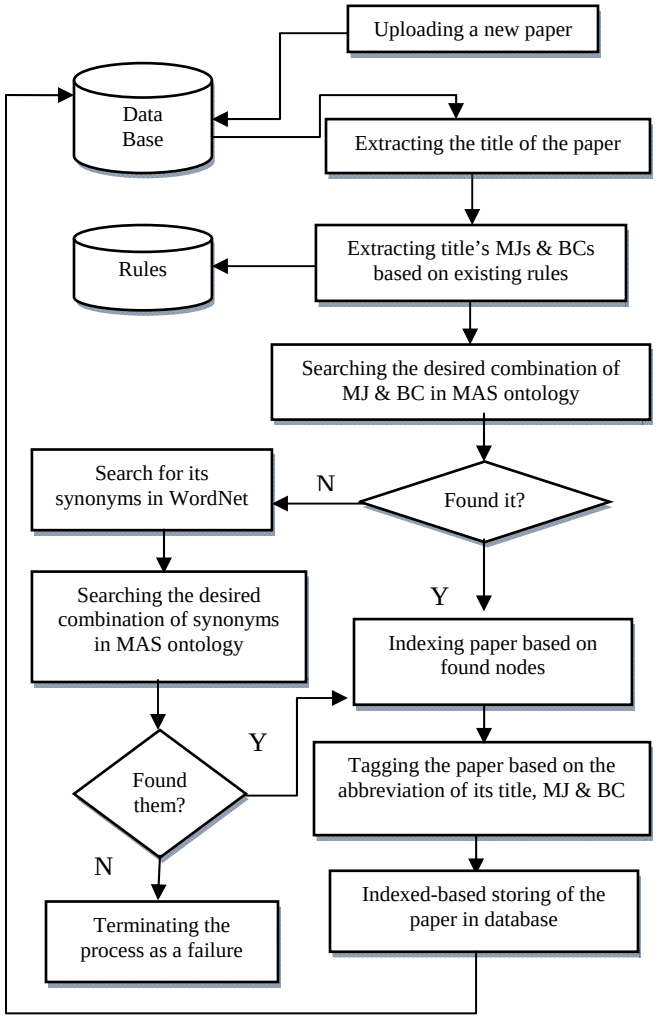

Figure 1. Flowchart of proposed indexing framework

Defining suitable rules as well as domain ontology can realize the indexing process appropriately. For this purpose, we are employing the ontology of agent science and technology that we have developed in Multi Agent Science Laboratory of University of Tehran for this purpose.

Searching the desired combination of major concern and basic constituent in agent science and technology ontology and finding the related nodes reveals the appropriate keywords for tagging and indexing the paper. It is to be noted that in cases where the major concern and basic constituent do not match any node of ontology, WordNet

seems to be a good realm for substituting alternative words. If the related words, in any of the mentioned ways be found, the paper will be tagged, indexed and respectively stored in the database. Otherwise, the process will be terminated as a failure. In the following sections, we briefly discuss each major constituent of proposed framework illustrated in Figure 1 including the functionality of major concern and basic constituent in indexing process. Followed by explaining the role of ontology and WordNet to support appropriate semantically indexing of papers and respectively storage of them.

B. The proposed indexing method based on major concerns and basic constituents

As previously mentioned, each title usually has two parts; major concern (MJ) and basic constituent (BC). Major concern is the part which explains about the main objective of the paper, while Basic Constituent mainly focuses on the methods, techniques or tools which were used to reach the objective in major concern [10, 12].

Reviewing several numbers of titles lead us to the following structure for MJ and BC. Four main parts are considered for MJ as follows:

1) Action part; which is mainly a verb.
2) Direct object; which is a noun or a pronoun that receives the action of a verb or shows the result of the action. It answers the question "What?" or "Whom?" after an action verb.
3) Indirect object; which is the recipient of the direct object and answers the question "To whom?" or "For whom?". It usually follows a preposition
4) Adverb/ Adjective part; which can modify verbs, adjectives, clauses, sentences, and other adverbs. It typically answers questions such as "how?", "in what way?", "when?", "where?", and "to what extent".

It has to be mentioned that some conjunctions like "in" and "for" followed by a verb usually yield into having two layers for MJ, which follow the same structure as mentioned above. Obviously, one or some of these parts may be absent in a title.

For BC, most of the time, maximum of one layer seems to be sufficient.

Having studied several titles, we gathered some rules, which were used to extract MJ and BC from a title. Certain conjunctions and prepositions can be signs of MJs and BCs. We prioritized some prepositions and conjunctions over others. Table 1 is a list of some of these prepositions and conjunctions.

TABLE I. LIST OF PRIORITIZED PREPOSITIONS AND CONJUNCTIONS

Priority	Preposition/ Conjunction
PR1 (BC)	Based on, on the basis of, on the ground of, using, making use of, taking into, ….
PR2 (MJ)	With the purpose of, with the aim of, in order to, in order that, …
PR3 (BC, MJ)	with, by, in, for, via, at, from, about, across, after, against, along, among, around, before, behind, beside, during, inside, instead of, onto, outside, over, since, through, under, within, …
PR4	and, of, into, like, without, both, together with, as, neither, either, as well as, rather than, than, …

Using the prepositions and conjunctions in the table above, we are able to detect BC and MJ before or after these words.

Employing a NLP parser can facilitate this process. For example, consider the title "Extending process automation systems with multi-agent techniques", as it is illustrated in the table, basic constituent can be found after "with", while before "with" we have major concern. It is to be noted that in MJ part, "extending" plays the role of action while the "process automation system" refers to direct object. Processing some complicated titles, necessitate more rules. For instance, "An agent-based signal processing in node environment for real-time human activity monitoring based on wireless body sensor networks" is a complicated title including several conjunctions and prepositions. Figure 2 illustrates MJ and BC of the title in detail.

MJ

	Action-part	Adverb-part	Direct Obj.	Indirect Obj.
MJ_1	**processing**	**agent-based**	**signal**	**Node environment**
MJ_2	**monitoring**	**Real time**	**Human activity**	-

BC

1st Layer :	2nd Layer
wireless body sensor networks	-

Figure 2. An example of major concern & basic constituent

Having reviewed large amount of titles, yield several rules for distinguishing MJs and BCs, Such as:

- If we find “via” or “based on” in the title, the following word or phrase will be BC.
- If we find “for” in the title, with a verb following, the rest of the title will be accounted as the 2nd layer of MJ.

C. *Ontology Processing*

After extracting MJ and BC based on the rules discussed in the previous section, combinations of MJ and BC are applied for seeking in the domain ontology. If the corresponding node is found, the keyword for annotating is achieved; otherwise closely related words or synonyms from WordNet have to be extracted for the same purpose. In this manner, indexing process based on the active nodes of ontology is realized. Vise versa, in the cases where not any related node is found, the process will be terminated and a failure notice will be issued. In situations where hierarchical ontology learning is considered, the new concept will be added in to the closest node of ontology. This would be our future trend of research in this subject.

Our domain ontology contains 200 nodes, with the depth of eight, describing agent science and technology. Figure 3 reveals a part of that ontology.

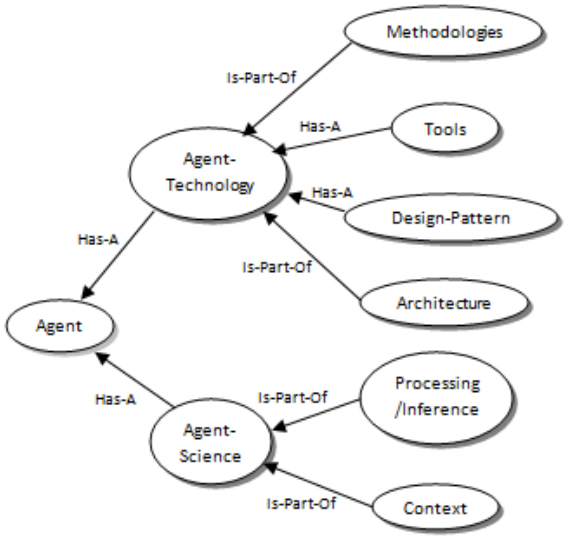

Figure 3. Part of the agent science & technology

IV. EXPERIMENTAL RESULTS

In order to evaluate the proposed approach, a data set of about 155 research papers in domain of agent science and technology which are collected from different conferences is considered.

We have indexed the research papers using the proposed framework. Figure 4 shows the pseudo code of the proposed framework.

Table 2 represents some examples of what our system produced for MJ, BC and related nodes in ontology. The papers which are shown in Table2 are as follows:

1: Group Communication based approach for Reliable Mobile Agent in information Retrieval Applications
2: Using a Dynamic Swarm of Intelligent Agents for Advising Farmers-AgroAgent
3: An Intelligent Inter Database Retrieval System Based on Multi-agent
4: The Personalized Information System of Lib2.0 Based on Agent

```
Extract Title;
  Parse (Title);
    Extract (MJ, BC);
      Search in Ontology (MJ, BC);
        Mark Related Nodes in Ontology;
          If (Marked Nodes == empty)
        Find Synonyms (MJ, BC);
      Search Synonyms in Ontology;
    Mark Related Nodes in Ontology;
Indexed-based Storage;
```

Figure 4. Psuedo code for proposed indexing framework

TABLE II. MJ, BC & ACTIVATED NODES IN ONTOLOGY FOR SOME PAPERS

Title	Major Concern (1st layer)	Major Concern (2st layer)	Basic Constituent	Activated nodes in Ontology
1	Group Communication based approach	Reliable Mobile Agent in information Retrieval Applications	-	Application, Mobile, Communication
2	Advising Farmers-Agro Agent	-	a Dynamic Swarm of Intelligent Agents	Dynamic, Reasoning
3	An Intelligent Inter Database Retrieval System	-	Multi-agent	Reasoning, Agent
4	The Personalized Information System of Lib2.0	-	Agent	-

As it is shown in table 2, our system has failed in finding an appropriate node for the 4th title because it wasn't able to find any of the words (and their synonyms) in our ontology. In the first two titles, our system has done very good in detecting MJ, BC and also in activating related nodes in ontology. For the 3rd title, despite having correctly extracting MJ and BC, our system mistakenly activated "Reasoning" node because of its similarity to the word "intelligent" which was in MJ of the title. For this reason, in future works, we have to apply better rules to avoid these mistakes.

Our system also shows great ability in finding MJ and BC of second layer. For example, for the title "Scalability and Load Balancing for Multiplatform Communication System Architecture based on Intelligent Agents", it gives "scalability and load balancing" as the MJ of the first layer and "Multiplatform Communication System Architecture" as the second layer of MJ, and therefore is able to detect "scalability, communication, architecture, reasoning" as the

 ISBN: 978-1-61208-181-6

corresponding nodes in the ontology. Overall, we were satisfied by the results our framework produced in processing MJ and BC for about 130 papers of the 155 papers.

In this paper, we used precision and recall measurements to judge the efficiency of our method. The "Precision" is calculated as the proportion of relevant retrieved documents to the number of retrieved documents and "Recall" is defined as the proportion of relevant retrieved documents to total number of relevant documents [26, 27].

$$\text{Precision} = TP/(TP + FP) \quad (1)$$
$$\text{Recall} = TP/(TP+FN)$$

where TP, TN, FP and FN denote true positives, true negatives, false positives, and false negatives, respectively.

Experimental results reveal that our system was able to index 118 number of papers correctly, while for 9 papers, our system issued a failure. It also made mistake in indexing 28 of the papers. Therefore, as it is illustrated in Figure 5, the

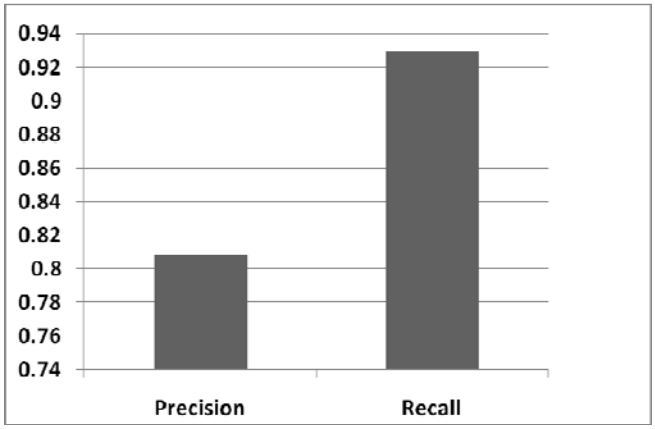

Figure 5. Precision and recall of experimental results

precision ratio is equal to 118/(118+28)= 0.81, while the recall ratio is 118/(118+9)=0.93.

In essence, our framework was able to correctly index 3/4 of the papers. The failures were mainly because of the following reasons: 1) Incompleteness of our ontology. 2) Lack of rules for extracting MJ and BC. 3) WordNet's inability to find scientific and agent-related words and phrases, and therefore not finding their synonyms. These problems can easily be resolved in future and as a result, the efficiency and correctness of our framework will be improved.

V. CONCLUSION AND FUTURE WORK

In this paper, we proposed an ontology-based indexing framework that can be used for managing the storage of papers in database. The proposed approach can also be an appropriate method for investigating the incremental learning of ontology. Distinguishing major concern and basic constituent of the title is our mean to find the corresponding nodes of ontology. To implement the proposed approach some NLP-based rules are proposed. It is mentioned in the paper that, in cases where there is no corresponding node in the ontology, WordNet is an appropriate supportive tool. Improving the rules from one perspective, and applying the proposed method for incremental ontology learning from another perspective are our future researches in this issue.

REFERENCES

[1] T. Hofmann, "Probabilistic latent semantic indexing," Proc. of the 22nd annual Intl. ACM SIGIR conference on Research and development in information retrieval, 1999, pp. 50-57.

[2] J. M. Gómez, J. C. Cortizo, E. Puertas, and M. Ruiz, "Concept Indexing for Automated Text Categorization," Natural Language Processing and Information Systems, Lecture Notes in Computer Science, Vol.136, 2004, pp. 495-502.

[3] J. Dobsa, " Comparison of information retrieval techniques: Latent semantic indexing (LSI) and Concept indexing (CI)," Solomonovi seminarji, Faculty of Organization and Informatics, Varazdin, University of Zagreb, Feb. 25, 2007.

[4] T. B. Lee, J. Hendler, and O. Lassila, "The semantic web," Scientific American, Vol. 284, No. 5, 2001, pp. 35–43.

[5] J. Kohler, S. Philippi, M. Specht, and A. Ruegg, "Ontology based text indexing and querying for the semantic web," Knowledge-Based Systems, Vol. 19, 2006, pp. 744–754.

[6] J. J. Chen and R. Sison, "Domain ontology learning," 5th Philippine Computing Science Congress, 2005, pp. 161-165.

[7] D. Sánchez, "Domain ontology learning from the web," The Knowledge Engineering Review, Vol. 24, Issue 4, 2009, pp. 413-413.

[8] M. Hazman, S. R. El-Beltagy, and A. Rafea, "A survey of ontology learning approaches," Intl. Journal of Computer Applications,Vol. 22, No.9, May 2011, pp. 36-43.

[9] L. Feigenbaum, M. N. Roy, and B. H. Szekely, W.C.Yung, "Ontology based text indexing," United States Patent Application 20080288442.

[10] K.Badie & M.T.Mahmoudi, " An approach to generating new ideas based on linking frames of concepts," ECCBR'04, Computational creativity workshop, 7th European Conf. In Case-based reasoning, Madrid, Spain, Aug 30- Sep 2, 2004.

[11] M.T.Mahmoudi & K.Badie, "Content determination for composite concepts via combining attributes' values of individual frames", IKE'04, Intl. Conf. On Information & Knowledge Engineering, Las Vegas, USA, Jun 21-24, 2004.

[12] K.Badie. M.T.Mahmoudi, "A Computational Framework for Manipulating an Issue from the View-Point of Other Issues," 14th Intl. Cong. of Cybernetics and Systems of WOSC – ICCS'08 ,Wroclaw, Poland, 2008, pp. 9 – 12.

[13] K.Badie, M.T.Mahmoudi, "View-Point Oriented Manipulation of Concepts: A Matching Perspective", IEEE Second International Conference on the Digital Society, ICDS 2008, Sainte Luce, Martinique, February 10-15, 2008.

[14] Y. He, S. C. Hui, and A. C. M. Fong, "Mining a web citation database for document clustering," Applied Artificial Intelligence Journal, Vol.16, 2002, pp. 283-302.

[15] K. M. Harnmouda and M. S. Kamel, "Phrase-based document similarity based on an index graph model," Second IEEE Intl. Conf. on Data Mining (ICDM'02), 2002, pp. 203-210.

[16] N. Bruno, L. Gravano, N. Koudas, D. Srivastava, "Navigation vs. index-based XML multi-query processing," 19th International Conference on Data Engineering (ICDE'03), 2003, pp. 139-150.

 ISBN: 978-1-61208-181-6

[17] D. D. Kha, M. Yoshikawa, and S. Uemura, "An XML indexing structure with relative region coordinate," 17th Intl. Conf. on Data Engineering, 2001, pp. 313-320.

[18] E. Dellis and G. Paliouras, "Management of large spatial ontology bases," Workshop on Ontologies-based techniques for DataBases and Information Systems (ODBIS) of the 32nd Intl. Conf. on Very Large Data Bases (VLDB 2006), 2006, pp. 102-118.

[19] S. T. Dumais, T. K. Landauer, and M. L. Littman, "Automatic cross-linguistic information retrieval using Latent Semantic Indexing," SIGIR'96, Workshop on Cross-Linguistic Information Retrieval, 1996, pp. 16-23.

[20] G. Karypis and E. Han, "Fast Supervised Dimensionality Reduction Algorithm with Applications to Document Categorization and Retrieval," 9th International Conference on Information and Knowledge Management (CIKM), 2000, pp. 12 – 19.

[21] M. Wan, A. Jonsson, C. Wang, L. Li, Y. Yang, "A random indexing approach for web user clustering and web prefetching," 2011 Workshop on Behavior Informatics, Shenzhen, China, 2011.

[22] N. Sugiura, K. Masaki, F. Naoki, I. Noriaki, and Y. Takahira, "A domain ontology engineering tool with general ontologies and text corpus," 2nd Workshop on Evaluation of Ontology based Tools, 2003.

[23] A. J. Yepes, R. B. Llavori, and D. R. Schuhmann, "Ontology refinement for improved information retrieval," Information Processing and Management, Vol. 46, 2010, pp. 426–435

[24] P. Martin and P. Eklund, "Embedding knowledge in Web documents," Computer Networks, Vol. 31, 1999, pp. 1403–1419.

[25] W. W. Chapman and J. N. Dowling, "Inductive creation of an annotation schema for manually indexing clinical conditions from emergency department reports," Journal of Biomedical Informatics, Vol. 39, 2006, pp. 196–208.

[26] D. L. Olson and D. Delen, "Advanced data mining techniques," Springer, 1st edition, February 1, 2008, pp. 138-138.

[27] C. D. Manning, P. Raghavan, and H. Schütze, "Introduction to information retrieval," Cambridge University Press., 2008.

 ISBN: 978-1-61208-181-6