American Scientific Research Journal for Engineering, Technology, and Sciences (ASRJETS)
ISSN (Print) 2313-4410, ISSN (Online) 2313-4402

http://asrjetsjournal.org/

Semantic Based Information Retrieval System Using Modified Inverse Document Frequency

Zun May Myint[a]*, Phyo Thuzar Tun[b]

[a,b]*Department of Computer Engineering and Information Technology Mandalay Technological University, Myanmar*

[a]*Email: ZunMayMyint@gmail.com*

[b]*Email: pinquinphyo@gmail.com*

Abstract

Today, Information Retrieval (IR) provides users with documents that will satisfy their information need. Word sense ambiguity is a cause of poor performance in IR system. IR performance will be increased if ambiguous words can be correctly disambiguated. Word Sense disambiguation (WSD) is the task to assign the correct meaning to such ambiguous words based on the surrounding context. Various senses provided by WSD process have been used as semantics for indexing the documents to aid the information retrieval system. K-Nearest Neighbour (KNN) is used for effective text classification in WSD process and the Vector Space Model (VSM) is used for IR process. The cosine similarity method is used in both KNN and VSM to calculate the similarity in which term frequency and inverse document frequency (TF-IDF) scheme is used to calculate the weight of each word. There is a challenge that the original TF-IDF scheme eliminates the related senses although there is a related sense. This paper thus proposes the modified TF-IDF method, so called TF-MIDF, to solve the no-relevant problem by modifying the IDF equation to improve the accuracy of IR performance. By comparing the performance between the original IDF scheme and the MIDF scheme, the average precision results of the original TF-IDF method is 71% and the average precision results of the TF-MIDF is 80%. Therefore, the proposed methodology is more precise than the original method while retrieving the relevant documents of the required information.

Keywords: Cosine Similarity; Modified Inverse Document Frequency (MIDF); IR; KNN Classifier; VSM; WSD; WordNet.

---

* Corresponding author.

1. Introduction

Nowadays, word sense ambiguity is a detrimental effect on the performance of text based information retrieval (IR) system [1]. Semantic indexing of the document changes from the keyword-based approach to the sense based approach for effective retrieval.

IR system will improve its performance if the documents it retrieves are represented by word senses rather than words. Overall, IR systems can potentially benefit from the correct meanings of words provided by WSD systems [2]. Word sense ambiguity can be thought of as the most serious problem in machine translation system .The human mind is able to select the proper target equivalent of any source language word by comprehension of the context. A human being may also automatically consider a group of words, rather than just one word, in order to understand the meaning of a sentence, even if the words of the group are not relevant[3]. As a basic semantic understanding task at the lexical level, WSD is a fundamental problem in natural language processing. It can be potentially used as a component in many applications, such as machine translation and information retrieval [4].

Word sense disambiguation method is needed for semantic indexing to get the correct sense of the indexed words. The semantic-based information retrieval system eliminates either the possibility of retrieving information that is obtained due to the presence of the irrelevant information that is retrieved because of no provision of the correct sense of the word in the searching process [5].

WSD approaches can be broadly classified into three categories. They are supervised approaches, semi-supervised approaches and unsupervised approaches. The supervised approaches use the machine-learning and data mining techniques to train a classifier from the sense-tagged corpora. The success of supervised learning approaches to word sense disambiguation is largely dependent on the features used to represent the context in which an ambiguous word occurs [6,7]. The unsupervised approaches do not use a training corpus and are based on the unlabelled corpora. The semi-supervised approaches are the hybrid of the two other categories. This system is based on KNN classifier in the supervised learning WSD method. KNN is a supervised learning approach in which the classification is accomplished by comparing a given test vector with training vector that are similar to it. When an unknown vector is introduced, the K-NN classifier finds k most similar training vectors that are closest to the unknown vector. These k training records are the k-nearest neighbor of the unknown vector. This classifier determines the label of the unknown vector by using its k nearest neighbors. This system provides additional semantics as conceptually related words with the help of glosses to each keyword in the query by disambiguating their meanings [8]. This system uses the WordNet and English corpus as the lexical resources that encode concepts of each term. WordNet is also used as the lexical resource to extract hypernyms, hyponyms, synonyms and gloss of ambiguous word [9] . In this sense, the cosine similarity method is used to choose the correct sense that is relevant to the ambiguous word in KNN and produce the relevant document in vector space model that is relevant to the user query by calculating the similarity [11]. In the cosine similarity, the Term Frequency and Inverse Document Frequency (TF-IDF) scheme is used for calculating the weight of each word. The original IDF eliminates the related senses.

So, modified TF-IDF scheme is proposed to solve no relevant problems for the semantic-based IR system. The proposed scheme is based on the K-Nearest Neighbor (K-NN) method in WSD approach and vector space model in IR process. So, this system provides the effectiveness of both KNN classifier based WSD method and semantic-based IR system with modified IDF.

The rest of the paper is organized as follows: Section 2 describes the explanation of the system with the original TF-IDF method, modified TF-IDF method and experimental results. Finally, conclusion is given in section 3.

2. Semantic Based Information Retrieval System

In the semantic based information retrieval system, the KNN classifier, cosine similarity method, TF-IDF weighting scheme, English corpus and WordNet are used for word sense disambiguation process. To search the similarity between the user query or the testing vector and the gloss of each sense or the training vector, this system uses the cosine similarity method and TF-IDF weighting scheme.

For retrieval process, the vector space model is used to calculate the similarity in which the cosine similarity method and TF-IDF weighting scheme is also used to calculate the similarity between the disambiguated query and documents. After converting each context to a vector of words, this system uses the cosine similarity method to measure the similarity between a new context and each existing context in the training corpus .

The cosine similarity between training vector d_j and testing vector q can be calculated as the following equation. Cosine similarity between two typical vectors can be defined as follow:

$$cosine(d_j, q) = \frac{\sum_{i=1}^{|v|} w_{ij} \times w_{iq}}{\sqrt{\sum_{i=1}^{|v|} (w_{ij})^2} \times \sqrt{\sum_{i=1}^{|v|} (w_{iq})^2}} \quad (1)$$

where, cosine(d_j, q) is cosine similarity between training vector d_j and testing vector q. W_{ij} is weight of the term t_i within training vector d_j. W_{iq} is weight of the term t_i within testing vector q [12].

A. Original Term Frequency and Inverse Document Frequency (TF-IDF) Weighting Scheme

To calculate the weight of each term in the training vector and testing vector, this system uses the TF-IDF weighting scheme. The term frequency is multiplied with the inverse document frequency to obtain the weight of each term. The term frequency within training vector is as follows:

$$tf_{ij} = \frac{f_{ij}}{max\{f_{1j}, f_{2j}, ..., f_{|v|j}\}} \quad (2)$$

where, f_{ij} is the raw frequency count of term t_i in training vector d_j and tf_{ij} is the normalize term frequency of term t_i in training vector d_j.

The original inverse document frequency (IDF) is as follow:

$$idf_i = log \frac{N}{df_i} \quad (3)$$

where, df_i is the number of train vectors in which term t_i appear. N is the total number of train vectors in the system. The idf_i is the inverse document frequency of term t_i.

The weight of the term within the training vector is as follow:

$$w_{ij} = tf_{ij} \times idf_i \quad (4)$$

where, w_{ij} is the weight of the term t_i in training vector dj.

The weight of the term within testing vector is as follows:

$$w_{iq} = \left[0.5 + \frac{0.5 f_{iq}}{max\{f_{1q}, f_{2q}, ..., f_{|v|q}\}} \right] \times log\left(\frac{N}{df_i} \right) \quad (5)$$

where, w_{iq} is the weight of the term t_i in testing vector q and f_{iq} is the raw frequency count of term t_i in the testing vector q [12].

B. Explanation of the proposed problem

The explanation of the system, the database consists of three documents as an example for the sample calculation. These documents are as follows:

Table 1: Sample document in the database

Document Name	**Information in the Document**
Document 1	DENCLUE framework builds on non-parametric method, namely kernel estimation.
Document 2	DENCLUE uses influence functions between data points to model the data space.
Document 3	DENCLUE method has strong mathematical foundation.

The user query is used for searching the user required information. The user query can be either ambiguous query or disambiguous query. The KNN classifier based on the WSD approach is used to disambiguate ambiguous query. Then, WordNet and English corpus are used as the lexical resources. As an example, the input query is as follows:

User Query: "Density Clustering"

Figure 1: Sample User Query

In the semantic information retrieval process, it is needed to perform the pre-processing process that is stopwords removal process. Example of stopwords are "is", "on", "with" and so on. In the above sample, the user query does not include stopwords. The "density" and "clustering" are keywords because these words are meaningful words.

After removing stopwords, each keyword is checked that is ambiguous or dismbiguous word. The senses from WordNet and English corpus are used for checking the ambiguous word or not. If the keyword has more than one sense, this word is assumed as ambiguous word. In the above sample, the "density" and "clustering" keywords are ambiguous word. Because of the "density" keyword has two senses from the WordNet and "clustering" keyword has two senses from the English corpus. Ambiguous word and its senses and glosses are shown in Table 2.

Table 2: Word senses and its gloss

Keyword	Sense ID	Sense Name	Gloss
density	Sense 1	denseness	The amount per unit size
	Sense 2	Concentration, denseness, tightness, compactness	The spatial property of being crowded together
clustering	Sense 1	DBSCAN	It is density based method that discovers clusters in spatial database.
	Sense 2	DENCLUE	It is density method. It is based on density distribution functions

To disambiguous the ambiguous word, the training and testing vectors are created about the ambiguous word. Training vectors are created by using the gloss of synonyms, hypernyms and hyponyms of each word from the WordNet and English corpus.

Testing vector is created by using the context in the user query. For each vectors, it is needed to remove

stopwords. Training and testing vector of each ambiguous word are shown in Table 3.

Table 3: Training vector and testing vector

Ambiguos Word	Vector Name	Vector
density	Training vector 1 for sense 1	[amount, per, unit, size]
	Training vector 2 for sense 2	[spatial, property, being, crowded]
	Testing vector	[clustering]
clustering	Training vector 1 for sense 1	[density, based, method, discovers, clusters, spatial, database]
	Training vector 2 for sense 2	[density, method, based, density, distribution, functions]
	Testing vector	[density]

The KNN classifier, cosine similarity and TF-IDF method are used to search the most relevant sense of each ambiguous word by classifying each training vector and testing vector. The calculated weight results by using TF-IDF method about “clustering” keyword are shown in Table 4.

The cosine similarity between training vector and testing vector is calculated to choose the most relevant sense of ambiguous word by using the weight result. According to the KNN classifier, to choose the most relevant one sense that has the highest similarity value among other similarity results, the K value is defined as K=1.

The most relevant sense of "clustering" cannot be retrieved by using the similarity result of the TF-IDF method as shown in Table 5. So, it produces the original query to the user as the disambiguated query.

In the semantic based information retrieval system, the required information is retrieved from the database by using disambiguous query.

For retrieval process, the cosine similarity method and TF-IDF weighting scheme is used to calculate the similarity between the disambiguated query and documents. As a sample, three documents in the database are shown in Table 1.

The TF, IDF and weight of each keyword are used to calculate the similarity between the user disambiguous query and each document. The weight results by using TF-IDF method is shown in Table 6.

After calculating weight of each keyword, these weights are used to calculate the cosine similarity.

These similarity results by using the original TF-IDF method is shown in Table 7.

Although there is relevant document in documents database, the original TF-IDF based similarity method cannot

retrieve the relevant document.

So, the system proposed the modified inverse document frequency (MIDF) scheme.

Table 4: Weight result about "clustering" keyword

Vector Name	Term (Keyword)	Weight Result from TF - IDF		
		TF-IDF		Weight
		TF	IDF	
Training Vector 1	density	1	0	0
	based	1	0	0
	method	1	0	0
	discovers	1	0.301	0.301
	clusters	1	0.301	0.301
	spatial	1	0.301	0.301
	database	1	0.301	0.301
Training Vector 2	density	1	0	0
	based	0.5	0	0
	method	0.5	0	0
	distribution	0.5	0.301	0.151
	functions	0.5	0.301	0.151
Testing	density	1	0	0

Table 5: Similarity result about "clustering" keyword

ID	Similarity between Training Vector and Testing Vector	Cosine Similarity Results using TF-IDF
1	Training Vector 1 and Testing Vector	0
2	Training Vector 2 and Testing Vector	0

The MIDF is proposed to retrieve more relevant senses, whereas original IDF does not retrieve it even though there is a relevant sense as explained in the above section. The proposed equation is shown as follow. In this equation, simpl y adding N/df_i to the original equation works perfectly for all input sentences while finding the relevant senses.

$$midf_i = log\left(\frac{N}{df_i} + \frac{N}{df_i}\right) \qquad (6)$$

where, df_i is the number of train vectors in which term t_i appear. N is the total number of train vectors in the system. idf_i is the inverse document frequency of term t_i.

Table 6: Weight result of each keyword

Name	Term (Keyword)	Weight Result from TF-IDF	
		TF-IDF	Weight
		TF IDF	
Document 1	DENCLUE	1 0	0
	framework	1 0.477	0.477
	builds	1 0.477	0.477
	non-parametric	1 0.477	0.477
	method	1 0.176	0.176
	namely	1 0.477	0.477
	kernel	1 0.477	0.477
	estimation	1 0.477	0.477
Document 2	DENCLUE	0.5 0	0
	uses	0.5 0.477	0.239
	influence	0.5 0.477	0.239
	functions	0.5 0.477	0.239
	data	1 0.477	0.477
	points	0.5 0.477	0.239
	model	0.5 0.477	0.239
	space	0.5 0.477	0.239
Document 3	DENCLUE	1 0	0
	method	1 0.176	0.176
	strong	1 0.477	0.477
	mathematical	1 0.477	0.477
	foundation	1 0.477	0.477
Query	DENCLUE	1 0.477	0.477
	density	1 0.477	0.477
	clustering	1 0.477	0.477

C. Explanation of the proposed solution

This system searches the most relevant sense of each ambiguous word and the relevant document between the disambiguated query and the documents by using the TF-MIDF weighting schemes. The same sample query is also used in the explanation of the TF-MIDF scheme. In WSD process, the KNN classifier, cosine similarity and TF-MIDF method are used to search the most relevant sense of each ambiguous word by classifying each

training vectors and testing vector. The calculated weight results by using TF-MIDF method about “clustering” keyword are shown in Table 8.

Table 7: Similarity result between disambiguous query and documents

ID	Similarity between Disambiguous Query and Each Document	Cosine Similarity Results using TF-IDF
1	Disambiguous Query and Document 1	0
2	Disambiguous Query and Document 2	0
3	Disambiguous Query and Document 3	0

Modified Inverse Document Frequency

Table 8: Weight result about “clustering” keyword

Vector Name	Term (Keyword)	Weight Result from TF - MIDF		
		TF-MIDF		Weight
		TF	MIDF	
Training Vector 1	density	1	0.301	0.301
	based	1	0.301	0.301
	method	1	0.301	0.301
	discovers	1	0.602	0.602
	clusters	1	0.602	0.602
	spatial	1	0.602	0.602
	database	1	0.602	0.602
Training Vector 2	density	1	0.301	0.301
	based	0.5	0.301	0.151
	method	0.5	0.301	0.151
	distribution	0.5	0.602	0.301
	functions	0.5	0.602	0.301
Testing	density	1	0.301	0.301

According to the similarity result by using TF-MIDF, the sense "DECLUE" about the training vector 2 is the most relevant sense of "clustering" ambiguous word. After disambiguation each ambiguous word from the user query, the disambiguous user query is as follows:

Disambiguous Query: DEClUE Density Clustering

Figure 2: Disambiguous Query

The similarity results by using the TF-MIDF method as shown in Table 9.

Table 9: Similarity result about "clustering" keyword

ID	Similarity between Training Vector and Testing Vector	Cosine Similarity Results using TF-MIDF
1	Training Vector 1 and Testing Vector	0.22942
2	Training Vector 2 and Testing Vector	0.53427

The cosine similarity method and TF-MIDF weighting scheme are used to calculate the similarity between the disambiguated query and documents in the retrieval process.

The TF, MIDF and weight of each keyword are used to calculate the similarity between the user disambiguous query and each document. The weight results by using TF-MIDF method is shown in Table 10.

After calculating weight of each keyword, these weights are used to calculate the cosine similarity. These similarity results by using the TF-MIDF method is shown in Table 11.

The document 3 is the most relevant to the disambiguous query according to the TF- MIDF based similarity results. So, the TF-MIDF based similarity method is more effective than TF-IDF based method. So, the modified IDF method is able to improve the performance of disambiguation process and IR process.

D. Experimental Result of the Proposed System

To access the "accuracy" or "correctness" of the system, there is a measure of IR success that is based on the concept of relevance. The effectiveness of an Information Retrieval system is evaluated by using precision.

Precision is the percentage of retrieved documents that is relevant to the query.

The precision can be defined as follows:

$$\text{Precision} = \frac{\text{Number of relevant documents retrieved}}{\text{Total number of documents retrieved}} \times 100\%$$

As a sample, the average precision results of the system are shown in Figure 3. These results are the accuracy results using TF-IDF and TF-MIDF based similarity method. In this sample, the relevance of result documents with user intended category is measured the number of relevant documents with user intended category by the total number of retrieved documents. As we assume, documents are ranked according to the scores of similarity, the top result documents are more close to user requirements. So, top 100 of retrieved documents are considered as the total number of retrieved documents.

Table 10: Weight result of each keyword

Name	Term (Keyword)	Weight Result from TF-MIDF		
		TF-MIDF		Weight
		TF	MIDF	
Document 1	DENCLUE	1	0.301	0.301
	framework	1	0.778	0.778
	builds	1	0.778	0.778
	non-parametric	1	0.778	0.778
	method	1	0.477	0.477
	namely	1	0.778	0.778
	kernel	1	0.778	0.778
	estimation	1	0.778	0.778
Document 2	DENCLUE	0.5	0.301	0.151
	uses	0.5	0.778	0.389
	influence	0.5	0.778	0.389
	functions	0.5	0.778	0.389
	data	1	0.778	0.778
	points	0.5	0.778	0.389
	model	0.5	0.778	0.389
	space	0.5	0.778	0.389
Document 3	DENCLUE	1	0.301	0.301
	method	1	0.477	0.477
	strong	1	0.778	0.778
	mathematical	1	0.778	0.778
	foundation	1	0.778	0.778
Query	DENCLUE	1	0.301	0.301
	density	1	0.778	0.778
	clustering	1	0.778	0.778

Table 11: Similarity result between disambiguous query and documents

ID	Similarity between Disambiguous Query and Each Document	Cosine Similarity Results using TF-MIDF
1	Disambiguous Query and Document 1	0.0399647
2	Disambiguous Query and Document 2	0.0321498
3	Disambiguous Query and Document 3	0.0543714

Table 12: Query and relevant documents

Query	Document	Retrieved	Relevant	Relevant-retrieved (TF-IDF)	Relevant-retrieved (TF-MIDF)
Clustering theory	100	22	17	16	16
Classification method based on tree logic	100	7	4	0	4
Density clustering	100	4	3	0	3
Classification system based on network logic	100	20	19	17	17
Outlier analysis in mining	100	20	19	19	19

As the sample, Given a query, the retrieved, relevant and then the relevant and retrieved documents for each query are shown in Table 12. The accuracy result using TF-IDF method is 71% and the accuracy result using TF-MIDF is 80%. The performance comparison result is shown in Figure 3.

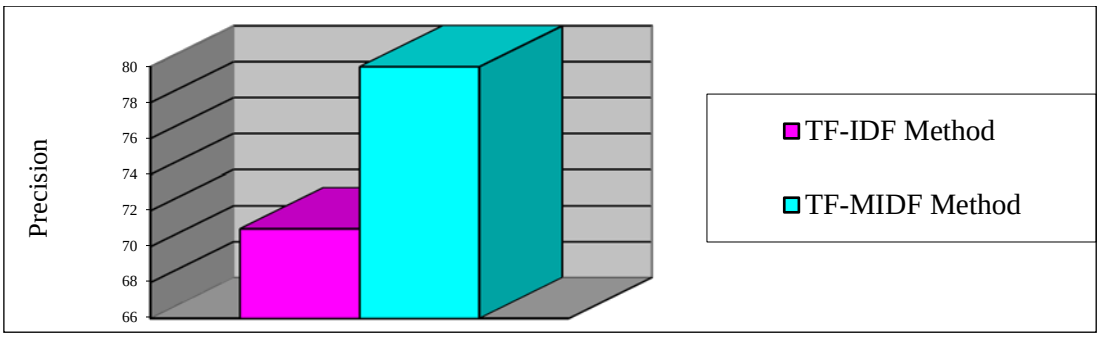

Figure 3: The average precision result of the system

3. Conclusion

Information Retrieval system will be better if ambiguous words can be correctly disambiguated. This paper proposes the modified IDF schemes of cosine similarity in KNN classifier and vector space model for the performance improvement of WSD and IR process. The original TF-IDF cannot retrieve the relevant senses and documents because this weighting scheme cannot search the similarity between training vector and testing vector. By increasing one or more factors, the results that more factors give higher similarity values. However, it cannot be assumed as better performance because the increasing factors of IDF give same disambiguated output query.

To know which weighting scheme is more effective for WSD process by increasing one or more factors of IDF equation, all modifications proved that and gave the same relevant sense and documents. Therefore, when the computation complexity and the processing time are considered, the variation of increasing factors that simply add N/df_i once to the original IDF equation is more effective for WSD and IR process. This system can be used in interesting application area by using specific domain such as data mining. Further research can be done to implement other domain areas of information retrieval system by adopting respective domain specific.

The proposed MIDF scheme which provides the better performance by simply adding N/df_i to avoid the problem of no-relevant sense and to retrieve the related documents. By comparing the performance between the original IDF scheme and the MIDF scheme, the average precision results of the original TF-IDF method is 71% and the average precision results of the TF-MIDF is 80%. Therefore, the performance of the semantic-based IR system using the proposed methodology gives a better precision.

Acknowledgements

The author especially would like to take this opportunity to express my sincere gratitude, respect and regards for supervisor Dr. Phyo Thuzar Tun, Assistance Lecturer, Department of Information Technology, Mandalay Technological University, under whose guidance, constant encouragement, patient and trust, I have worked on this paper. And the author is thankful to all teachers in Department of Information Technology, Mandalay Technological University, for their effective guidance, helpful suggestion and supervision for this paper and all of my friends who have directly or indirectly assisted me in my endeavors.

References

[1] S. Christopher and P.Oakes, " Word Sense Disambiguation in Information Retrieval Revisted", The University of Sunderland , Informatic Centre, August, 2003, Canada.

[2] D. Subarani, “Concept Based Information Retrieval from Text Documents”, Dept. of Computer Sciences, SLN College of Sciences, Tirupathi, India, IOSR Journal of Computer Engineering (IOSRJCE), PP 38-38, July-Aug, 2012.

[3] S. Viswanadha Raju, J. Sreedhar and P. Pavan Kumar, “Word Sense Disambiguation: An Empirical

Survey", International Journal of Soft Computing and Engineering (IJSCE), Volume-2, Issue-2, May, 2012.

[4] R. Navigli, "Word Sense Disambiguation: A Survey", ACM Computing Surveys, Vol. 41, No. 2, Article 10, Italy, February, 2009.

[5] N. Sharma and S. Niranjan, " An Optimized Combinational Approach of Learning Algorithm for Word Sense Disambiguation", International Journal of Science and Research (IJSR), vol 3 Issuse 6, June 2014.

[6] P. Tamilselvi and S. K. Srivatsa (2011), "Word Sense Disambiguation using Case based Approach with Minimal Features Set", Indian Journal of Computer Science and Engineering (IJCSE), vol. 2, no. 4, pp. 628-633, 2011.

[7] P. Tamilselvi and S.K. Srivatsa , "Case Based Word Sense Disambiguation Using Optimal Features", International Conference on Information communication and Management IPCSIT vol. 16, 2011, Singapore

[8] A. R. Rezapour, S. M. Fakhrahmad and M. H. Sadreddini, "Applying Weighted KNN to Word Sense Disabiguation", Proceedings of the World Congress on Engineering, Vol III, U.K, July 6-8, 2011.

[9] M. Barathi and S. Valli , "Ontology Based Query Expansion Using Word Sense Disambiguation", International Journal of Computer Science and Information Security(IJCSI), vol. 7. No.2, February 2010.

[10] Donald Metzler "Generalized Inverse Document Frequency", Napa Valley, California, USA, October 26-30, 2008.

[11] M. Nameh, S.M. Fakhrahmad and M. Zolghadri Jahromi , "A New Approach to Word Sense Disambiguation Based on Context Similarity", Proceedings of the World Congress on Engineering(WCE), vol 1. July 6-8, 2011, London, U.K.,

[12] B. Liu, Web Data Mining, Department of Computer Science, University of Illinois at Chicago, USA, 2007.