International Journal of Applied Engineering Research ISSN 0973-4562 Volume 11, Number 15 (2016) pp 8708-8713

Sense-Based Information Retrieval Using Artificial Bee Colony Approach

Dr. Alia Karim Abdul Hassan[1] and Mustafa Jasim Hadi[2]
[1,2]College: Computer Science Department, University of Technology/ Baghdad,
[1,2]Postal address: Al Rusafa-10045,
[1,2]City, state, zip code: Baghdad, Alsenaastreet, 10,
[1,2]Country name: Iraq,
e-mail: [1]hassanalia2000@yahoo.com,[2] mustafa_awadi@yahoo.com

Abstract
The combination of Information Retrieval (IR) and Word Sense Disambiguation (WSD) is still a big challenge in addressing the large-scale databases and the Web. The traditional search by keywords in IR systems has two problems. The first is the ranking to some of results that don't satisfy the user's need. The second is the potential similarity of documents in the ranked results. This makes users spend much time to organize and find the satisfied search results. Another significant problems are the huge information, in the large-scale databases and especially the Web, that lead to an exponential increasing at the searching time, and also the improving of the retrieval quality that is still the subject of controversy among a lot of authors. So, this work aims to develop an innovative model using bio-inspired approach called Artificial Bee Colony (ABC) to improve IR with consider the WSD problem. The improving is measured by the latency and solution quality. WSD problem is solved by using the Simplified Lesk algorithm.

Keywords: Information Retrieval; Word sense Disambiguation; Artificial Bee Colony; Simplified Lesk Algorithm.

INTRODUCTION

Information retrieval (IR) is an important field of computer science concerned with the storage, search and retrieve the information such as texts, Web pages,images,and videos.It is heavily dependent on Natural Language Processing (NLP), such as the stop word removal and the word stemming or lemmatization. Recently, a high-level process called Word Sense Disambiguation (WSD) has become one of the central challenges in the NLP field. The WSD aims to find the correct meaning (sense) of a word in a context [1].

With the huge information as in the large-scale databases and the Web, the traditional algorithms are incompetent in dealing with user queries in real time, i.e. they lead to an exponential increasing at the searching time. Bio-inspired approaches and more precisely swarm intelligence approaches can be used to get an efficient search for the huge information that find the information needed in an interesting record time [2], [3], [4].

Swarm intelligence is an area that emerged to serve the optimization field. The term "swarm" refers to a collection of animals such as birds, fishes, ants, bees, and termites [5]. Artificial bee colony (ABC) optimization algorithm is a recently introduced population-based and meta-heuristic approach formulated on the basis the natural behavior of bees in search for food [6].

This work uses the ABC algorithm after significant modifications. The motivation is to overcome challenges in using the WSD with IR together in a single system to retrieve only the most accurate and relevant documents that satisfy the user's need. The work presents a new approach inspired from ABC algorithm called WSD-MABC that includes WSD in the searching mechanism. In IR, the major challenge is related to three measures mentioned in IR literatures are the precision, recall, and the response time or the latency, while the major challenge in WSD is the finding out the correct sense of each word in its context. In this work, WSD problem is solved by using the Simplified Lesk algorithm and the lexical resource that is used is the WordNet dictionary. This dictionary is an online lexical resource developed at Princeton University.

WORD SENSE DISAMBIGUATION

Natural languages, in general, include a lot words have different meanings in different contexts. For example, the word *bank* in English may have various meanings such as *financial institution*, *reservoir*, *river side* etc. The Word sense disambiguation (WSD) is a mechanism used to search the correct meaning of a word which have multiple meanings, this is done by taking into account the location of the word in the context [7]. There are several available methods have been employed to address the ambiguous words, among them the most common method called "Lesk algorithm".

A simplified version of the Lesk algorithm was described in [8] and is often called the "Simplified Lesk algorithm". The basic idea of the Simplified Lesk algorithm is to disambiguate each word in a phrase separately. Given a word, the meaning is selected so that its gloss or definition offers the maximum similarity with its context [9]. Lesk algorithm (includes the Simplified Lesk algorithm) is a dictionary-based algorithm and the version that uses Word Net is reported to get good WSD results [10].

Word Net has been developed at Princeton University and is a lexical database system used online. This system models the lexical knowledge of English native speaker. It consists of nearly 100,000 terms classified hierarchically. Noun, verb, adjective and adverb for each word are grouped into synonym sets. The synonym sets of all words are also organized into senses (meanings). Synonym sets can also be related to Hyponym/Hypernym, and the Meronym/Holonym [11].

INFORMATION RETRIEVAL

Information retrieval (IR) is a process to represent, store, organize and search the information items. Information must

be structured in some manner that ensures the relevant information retrieval [12]. There are four basic components of this process are [2]:

1- The document that can be a text, Web page, image, or video.
2- The query that represents the user need.
3- The similarity formula that computes the similarity between a query and a document.
4- The test evaluations such as the precision and recall. Precision is "*the fraction of retrieved documents that are relevant*" and the recall is "*the fraction of relevant documents that are retrieved*".

Many IR models are mentioned in the literature. Vector space model (VSM) is the most widely used in which documents and queries are stored in weights vectors.

The weight vector of a document d will form $<w_{d,1},w_{d,2},...,w_{d,n}>$, while the weight vector of a query q will form $<w_{q,1},w_{q,2},...,w_{q,n}>$, the inner product between these two vectors is computed to find the similarity between a query q and a document d as follows:

$$Similarity\ (q,d) = \sum_{t=1}^{n} w_{q,t}.w_{d,t} \quad (1)$$

The computed similarity is important to rank, in descending order, the retrieved documents. Best fully weighting system, which is a typical term-weighting scheme, uses weights for the document terms formed by a cosine normalized $tf * idf$ and weights for the query terms formed by an enhanced but not normalized $tf * idf$. The document term weight $w_{d,t}$ is computed as follows:

$$w_{d,t} = \frac{tf_{d,t} \times log\ (N/df_t)}{\sqrt{\sum_{vector}\left(tf_{d,t} \times log\ (N/df_t)\right)^2}} \quad (2)$$

And the query term weight $w_{q,t}$ is computed as follows:

$$w_{q,t} = \left(0.5 + 0.5\ \times \frac{tf_{q,t}}{max\ tf_{q,t}}\right) \times log\ (N/df_t) \quad (3)$$

Where tf stand for *term frequency* and the expression $log(N/df)$ refers to idf that stands for *inverse document frequency*. *N* is the number of documents in a document collection and df is the number of documents where the term appears [13].

The most common traditional search approach in IR applications is inverted-index based search. Through using the inverted index, the search complexity of query-document with non-zero similarity is reduced at aphenomenal rate. However, the inverted index file may become untreatable for the large-scale databases or the web. Metaheuristic approaches such as the swarm intelligence can get a response time with a polynomial rate on a higher computation scale [2], [4].

ARTIFICIAL BEE COLONYALGORITHM

Artificial Bee Colony (ABC) algorithm is a stochastic-based poi-inspired algorithm falls under the umbrella of swarm intelligence. ABC is introduced by Karaboga [14] and described in a general form in [15] as follows:

- Initialization Phase
- REPEAT
  - ✓ *Employed Bees Phase*
  - ✓ *Onlooker Bees Phase*
  - ✓ *Scout Bees Phase*
  - ✓ *Memorize the best solution achieved so far*
- UNTIL requirements are met.

In the initialization phase, the initial solutions are computed as follows:

$$x_{ij} = x_{minj} + rand\ [0, 1] * (x_{maxj} - x_{minj}) \quad (4)$$

Where:

$i \in \{1,..,N_S\}$, $j \in \{1,..,D\}$, N_S is the number of food sources, and D is the number of optimized parameters,

x_{maxj} and x_{minj} refere to the upper and lower bounds for the dimension j.

*rand[0, 1]*is a random number between [0, 1].

In the employed bees and onlooker bees phases, the new solutions v_{ij} are computed as follows:

$$v_{ij} = x_{ij} + \phi_{ij}\ (x_{ij} - x_{kj}\ ) \quad (5)$$

Where:

$i,k \in \{1,..,N_S\}$ and $j \in \{1,..,D\}$ are randomly chosen indexes,

$\phi_{ij}$is a random number between [-1, 1].

The selection of the new solutions in the onlooker bees is depending on probability $p_i$that is computed as follows:

$$p_i = 0.9 \times \frac{f_i}{f_{max}} + 0.1 \quad (6)$$

Where:

f_i is the fitness value of the food source *i*,

f_{max} is the maximum fitness of food sources.

In the Scout bees phase, a new source v_{ij} is randomly generated instead of an abandoned one depending on the equation (4).

RELATED WORKS

Ramya and Shreedhara [16] presented in their paper a brief review on the application of swarm intelligence to information retrieval with focused on the large-scale databases and the Web. They present several approaches and there are two among them have been the most related:

1- *Habiba Drias, Hadia Mosteghanemi* [2] attempted to improve the retrieving quality and addressing the time consumption by the information retrieval systems in an environment with massive information a and especially the Web. For this task they designed a Bees Swarm Optimization (BSO) algorithm to get into the prohibited documents and obtain the most relevant results. They used CACM and RCV1 corpuses to test their experiments.
2- *Hasanen S. Abdullah and Mustafa J. Hadi* [17] used another version of bees swarm intelligence called Artificial Bee Colony (ABC) algorithm that had been modified by adding augmented data structure in order to improve the neighbors search. They compared their work with the traditional inverted index approach. CACM and NPL corpuses were used to test their experiments.

Although the works described above improve response time and the solution quality, but they were based only on keywords to compute the similarity between queries and documents and lack to consider the sense of the word in its text. Finding out the correct sense of a word in a text for improving the information retrieval is still a challenge factor depending on sayings of many authors.

Boshra F. ZoponAl_Bayaty and Shashank Joshi [1] presented a literature review about the combination between WSD and IR. They list some of researches and studies about the using of WSD with IR. Some of the presented works reveal that there is a controversy is evident in the possibility of using the WSD with IR. Among the most prominent of these works is for the author *Hwee Tou Ng* in [18], who presented a review about the controversy in the ability of the WSD to improve IR. The author exhibits grouping of authors who are answer by yes or no for aquestion "Does WSD helps in IR?".

PROPOSED SENSE-BASED IR SYSTEM

Although the authors who support the trend that WSD can improve IR but they are not focused on the delay that WSD can cause it during the testing for all ambiguous words, in addition to the original time of IR process.

The proposed system tries to address this significant delay and taking into account the improving in the retrieval quality. This is done using an algorithm inspired from the swarm intelligence optimization field. The WSD-MABC algorithm is the search approach used in this work. WSD-MABC is inspired form ABC algorithm that is described in advance section. The proposed system with WSD-MABC algorithm is described in Fig. (1) below:

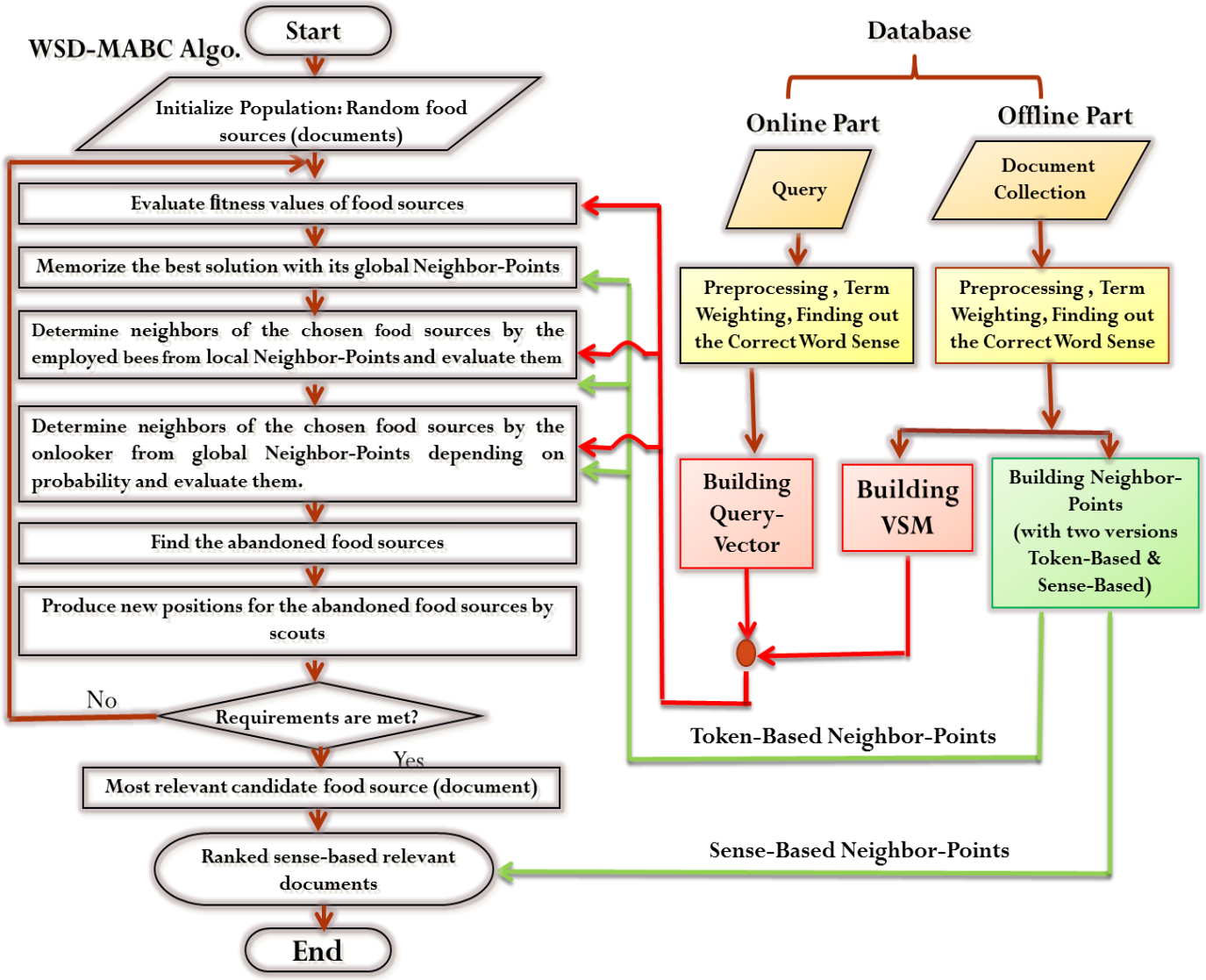

Figure 1: shows the proposed system.

The main factor in the success of the WSD-MABC algorithm described in the Fig. (1) is the using of neighboring points data structure. This data structure is offline achieved after the Vector Space Model (VSM) has been completed. Neighbor-points data structure is simply a ranked retrieving result associated with each document in the collection as if that

document is a query. Two versions of neighboring points feed the WSD-MABC algorithm. The first version called "*Token-Based Neighbor-Points*". It is a set of a sub-collection of documents, each sub-collection is clustered around a central document, but doesn't consider the word sense. Also, it relies on traditional weights and the similarity formula in the equation (1). Local Neighbor-Points referes to the sub-collection that is clustered around the current central document, while the global Neighbor-Points referes to the sub-collection that is clustered around the best so far central document. The second version, called "*Sense-Based Neighbor-Points*" is built just as "*Token-Based Neighbor-Points*" but it considers the correct word sense processing.

There are two phases of WSD-MABC algorithm. The first depends on the first version, with label *Token-Based Neighbor-Points*. The main idea is to get first the best relevant document without the consideration to the senses of the words. So, the system ensures that a specific document is the most relevant to the query without the side effects for matching of the senses that may in certain cases lead to a drop in precision and recall. The second phase of WSD-MABC depends on the second version, with label *Sense-Based Neighbor-Points*. It produces a group of the most relevant documents with the consideration of matching of the senses. The idea is first to obtain the best neighboring documents (from the *Sense-Based Neighbor-Points*) that are associated with the best relevant document (that is obtained from the first phase), and then to compute the query-documents similarity with consider the correct word sense.This work modifies the similarity equation in the equation (1) to a new similarity equation as follows:

$$Similarity(q,d)=\sum_{t=1}^{n} Q_t * D_t \quad (7)$$

Where Q_t and D_t are computed as follows:

$$Q_t = \left(0.8 * w_{q,t} + 0.2 * \propto\right)$$

$$D_t = \left(0.8 * w_{d,t} + 0.2 * \beta\right)$$

The parameters $\propto$ and β are computed as follows:

$$\propto = \gamma + \delta, \beta = \gamma + \omega$$

The parametersγ,δ, and ω are computed as follows:

$$\gamma = \left(\frac{No.ofmatchingsenses}{No.ofmatchingtokens}\right)$$

$$\delta = \left(\frac{No.ofmatchingtokens}{Totalno.oftokensinthequery}\right)$$

$$\omega = \left(\frac{No.ofmatchingtokens}{Totalno.oftokensinthedocument}\right)$$

For each matching word, the system checks whether the sense is also matching, and if otherwise it will ignore that word by make its similarity equals zero.In case the sense is matched, the system will compute the similarity using equation (7).

EXPERIMENTAL RESULTS

The proposed system is experimented on two different corpuses CACM (3204 documents, 64 queries with their relevance judgments) and NPL (11429 documents, 93 queries with their relevance judgments). These corpuses are well-known and used in many research works for evaluating IR systems. The proposed system is run on a personal computer (Core-i5 @2.50 GHz, RAM 6GB, 64-bit operating system).The experimental evaluations focus on the comparison between the proposed system and the traditional inverted-index system in two versions, with and without using WSD mechanism. To highlight the comparison, a sample of five queries is selected from each corpus with the aim to get different solutions that raises the controversy in the possibility of using the WSD with IR. Table (1) shows the relevant documents within rank 10 to a sample of CACM queries (q7, q21, q23, q48, q58) using both the traditional (with the two versions) and proposed systems. Table (2) shows the average of the performance evaluation of the sample queries in the CACM collection.

Fig. (2) shows the 11-point interpolated recall-precision curves for the different algorithms. The curves constructed using the average of precision and recall at rank 10 of the sample queries.For the NPL collection, tables (3) and (4) and the Fig. (3) are presented to show the performance with respect to a sample of NPL queries (q17, q19, q22, q61, q83).

Table 1: The relevant documents for a sample of CACM queries using the traditional (with and without WSD) and proposed algorithms.

Algorithms	Relevant docs for query q7	Relevant docs for query q21	Relevant docs for query q23	Relevant docs for query q48	Relevant docs for query q58
Traditional without WSD	[2912, 2376, 3043, 3148]	[1429]	[]	[2325]	[1344, 2249, 2634, 1709, 2098, 1944, 1631]
Traditional with WSD	[2912, 2376, 3043, 2256, 3128]	[2932, 2703]	[]	[]	[1344, 2249, 2634, 1709, 2098, 1398]
WSD-MABC	[2912, 3043, 2376, 2865, 2866, 3128]	[3018, 2932, 2702, 2703]	[]	[1353]	[2634, 2098, 1709, 1398]

Table 2: Average performance evaluation with the CACM collection of the sample queries

Average performance for CACM collection	Traditional without WSD	Traditional with WSD	WSD-MABC
The average of the documents that have been visited for each query out of (3204) documents	2394	2394	1446
Total no. of non-matching senses out of all the matching words	0	5514 out of (19336)	3285 out of(12063)
Averageof latency (Sec. /Query)	1.113672	1.298409	1.012170
Averageof precision	0.260000	0.260000	0.300000
Averageof recall	0.110087	0.112078	0.158918

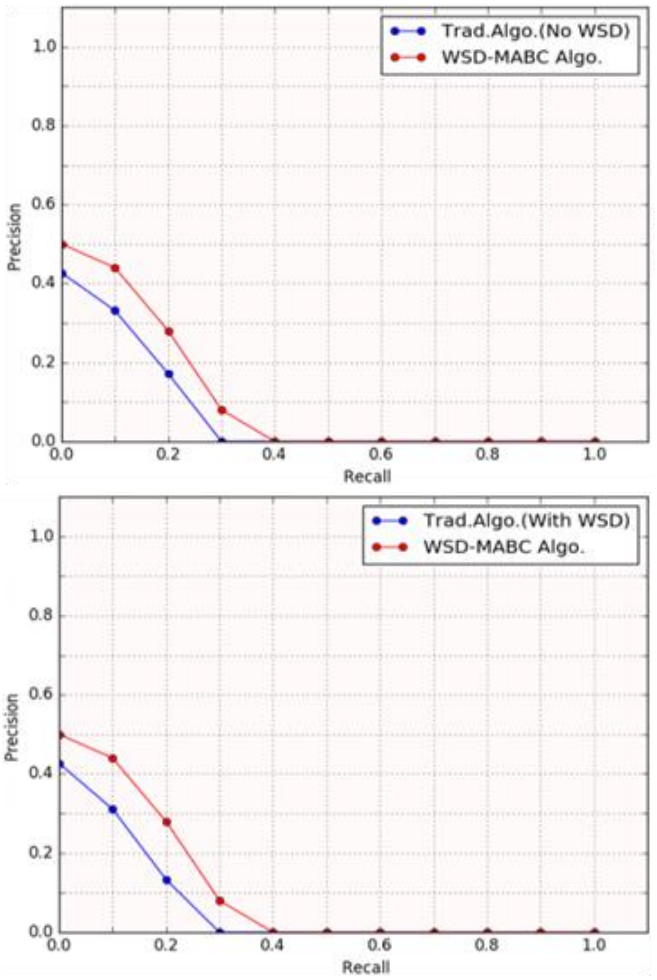

Figure 2: Average recall-precision curves for a sample of five CACM queries

Table 3: The relevant documents for a sample of NPL queries using the traditional (with and without WSD) and proposed algorithms.

Algorithms	Relevant docs for query q17	Relevant docs for query q19	Relevant docs for query q22	Relevant docs for query q61	Relevant docs for query q83
Traditional without WSD	[496, 11128, 10981, 9486, 4151]	[10693]	[608]	[2280, 7065, 9054, 8294]	[4629]
Traditional with WSD	[496, 10981, 9486]	[10693, 2895]	[608, 6026, 6403, 6027, 94]	[2280]	[]
WSD-MABC	[496, 10981, 9486]	[8757, 7615, 10693, 9053]	[11394, 6027, 6221, 3446, 91, 94, 9247]	[8768, 2280, 7442, 1996]	[3641, 6995, 1487]

Table 4: Average performance evaluation with the NPL collection of the sample queries

Average performance for NPL collection	Traditional without WSD	Traditional with WSD	WSD-MABC
The average of the documents that have been visited for each query out of (11429) documents	3598	3598	2088
Total no. of non-matching senses out of all the matching words	0	11167out of(24554)	1779 outof (9322)
Averageof latency (Sec. /Query)	1.366360	1.344202	0.961844
Averageof precision	0.240000	0.220000	0.420000
Averageof recall	0.081628	0.062900	0.123083

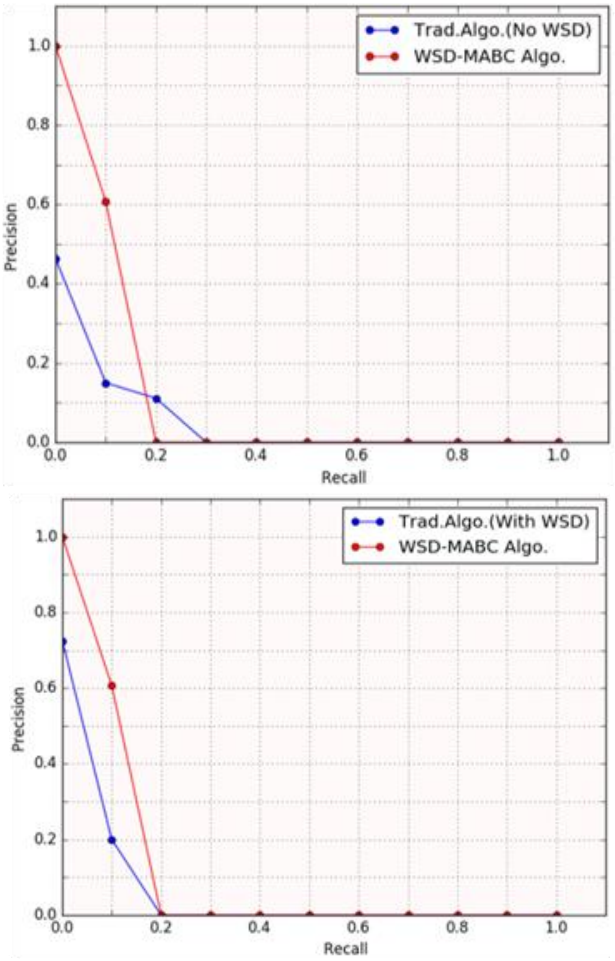

Figure 3: Average recall-precision curves for a sample of five NPL queries

DISCUSSION

Although the relevant documents presented in tables (1) and (3) exhibit a low precision of the proposed system in some queries, for example query q58 in CACM collection and query q17 NPL collection. However, the majority of the

queries gets results with equal or higher precision and recall in comparison to the two versions of the traditional algorithm. This indicates that the proposed system is a good solution to the controversy and discussion on the potential success of the use of the WSD with IR. In general, the numerical results in Table (2) and Table (4) exhibit the superiority of the proposed system on the traditional system in terms of the efficiency and effectiveness. Also the figures (2) and (3) show the superiority of the proposed system in the precision and recall with respect to the preceding appearing of the documents to the user.

CONCLUSIONS

In this paper, we developed a tool uses the Artificial Bee Colony (ABC) algorithm with modifications for improving the performance of the sense-based IR systems. The proposed system has overcome two important issues. The first is that the response time is very high in the systems that use traditional algorithms.The second is the improvement of retrieval quality which is the subject of controversy among the other authors.

The experimental results exhibit the superiority of the proposed system in terms of the precision, recall and latency in comparison to two versions of the traditional algorithm (with and without using WSD).

FUTURE WORK

The proposed system uses Simplified Lesk algorithm as a traditional disambiguation algorithm. In future work, an innovated is ambiguation tool inspired form the stochastic optimization algorithms can be used to increase the disambiguation accuracy. This will follow an increasing in the precision and recall of the sense-based IR. However, this should not be on the account of time consuming in the query processing stage and the basic issue lies in this task.

REFERENCES:

[1] Al_Bayaty, B.F.Z. and Josh, S.,2014," Word Sense Disambiguation (WSD) and Information Retrieval (IR): Literature Review", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 2.

[2] Drias, H. and Mosteghanemi, H.,2010, "Bees Swarm Optimization based Approach for Web Information Retrieval", IEEE/WIC/ ACM International Conference on Web Intelligence and Intelligent Agent Technology.

[3] Drias, H.,2011, "Web Information Retrieval using Particle Swarm Optimization based Approaches", IEEE/ WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology.

[4] Drias, H.,2011, "Parallel Swarm Optimization for Web InformationRetrieval", IEEEThird World Congress on Nature and Biologically Inspired Computing.

[5] Karaboga, D. and Akay, B.,2009,"A survey: algorithms simulating bee swarm intelligence",ArtifIntell Rev (2009) 31: 61-85DOI 10.1007/s10462-009-9127-4, Springer Science+Business Media B.V.

[6] Abu-Mouti, F. S. and El-Hawary, M. E.,2012," Overview of Artificial Bee Colony (ABC) algorithm and its applications", IEEE International, Systems Conference (SysCon), pp. 1-6.

[7] Pal, A.R.andSaha, D.,2015," Word sense disambiguation: A survey" International Journal of Control Theory and Computer Modeling (IJCTCM) Vol.5, No.3.

[8] Kilgarriff, A. and Rosenzweig,j.,2000,"Framework andresults for English SENSEVAL". Computers and the Humanities, 34, 15-48.

[9] Basile, P., Caputo, A., and Semeraro, G.,2014,"An Enhanced Lesk Word Sense Disambiguation algorithm through a Distributional Semantic Model"Proceedings of COLING, the 25th International Conference on Computational Linguistics: Technical Papers, pages 1591-1600.

[10] Shallu and Gupta, V.,2013," A Survey of Word-sense Disambiguation", Effective Techniques and Methods for Indian Languages // Journal of French and Francophone Philosophy, № 4 (5). C. 354-360.

[11] Varelas, G., Voutsakis, E., Petrakis, E. G. M.,Milios, E. E., and Raftopoulou, P.,2005," Semantic Similarity Methods in WordNet and their Application to Information Retrieval on the Web", 7 th ACM Intern. Workshop on Web Information and Data Management.

[12] Baeza-Yates, R. and Ribiero-Neto, B.,1999, "Modern Information Retrieval," Addison Wesley Longman Publishing Co. Inc.

[13] Kang,H. and Choi, K., 1997," Two-Level Document Ranking Using Mutual Information in Natural Language Information Retrieval", Information Processing and Management, vol. 33, no. 3, pp. 289-306.

[14] Karaboga, D., 2005," An idea based on honey bee swarm for numerical optimization".Technical Report TR06, Computer Engineering, Department, Erciyes University, Turkey.

[15] Karaboga, D.,Gorkemli, B.,Ozturk, C. andKaraboga,N., 2014," A comprehensive survey: artificial bee colony (ABC) algorithm and applications", Artificial Intelligence Review; 42(1), pp.21-57, DOI: 10.1007/s10462-012-9328-0.

[16] Ramya,.C and Shreedhara, K. S,2016,"A Brief Review On The Application Of SwarmIntelligence To Web Information Retrieval",International Journal of Engineering Research in Computer Science and Engineering (IJERCSE) Vol 3, Issue 1.

[17] Abdullah, H. S. and Hadi, M. J.,2014,"Artificial Bee Colonybased Approach for WebInformation Retrieval", Eng. & Tech. Journal, vol.32, Part (B),No. 5, pp. 899-909.

[18] Tou Ng, H., 2011, "Does Word Sense DisambiguationImprove Information Retrieval?", ESAIR'11, Glasgow, Scotland, UK.ACM 978-1-4503-0958-5/11/10.