Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 46 (2015) 329 – 336

Procedia Computer Science

International Conference on Information and Communication Technologies (ICICT 2014)

Ontology-Based Document Mining System for IT Support Service

Niloofer Shanavas[a,]*, Shimmi Asokan[a]

[a]*Department of Computer Science & Engineering, Rajagiri School of Engineering & Technology, Kochi - 682039, Kerala, India*

Abstract

Information Technology (IT) is a vital and an integral part of every organization. IT executives are constantly faced with problems that are difficult to tackle and time consuming. Experience is required to solve these problems easier and faster. We can utilize case-based reasoning (CBR), data mining and information retrieval (IR) techniques to automate IT problem solving and experience management. In this paper, we propose an IT ontology-based system for semantic retrieval that increases the efficiency and quality of IT support service. The proposed approach retrieves similar problem/solution pairs based on the concepts in the query and performs better than the traditional keyword-based approach especially in cases where the keywords of the relevant documents do not match the keywords in the query.

© 2015 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).
Peer-review under responsibility of organizing committee of the International Conference on Information and Communication Technologies (ICICT 2014)

Keywords: Semantic Retrieval; Case-based reasoning; Data Mining; IT problem management; IT support; Experience management.

1. Introduction

Today, Information Technology is a vital part of an organization and plays a very important role in its functioning. The quality and efficiency of IT support service largely depends on the experience of IT executives to solve problems. IT problem solving requires experience for quick problem resolution. Collecting, sharing and reusing experience helps to correct problems faster, improves the efficiency of IT support service and minimizes interruptions in the functioning of an organization. Case-based reasoning, data mining and information retrieval

* Corresponding author. Tel.: +91-8281235304.
E-mail address: niloofershanavas@hotmail.com

1877-0509 © 2015 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).
Peer-review under responsibility of organizing committee of the International Conference on Information and Communication Technologies (ICICT 2014)
doi:10.1016/j.procs.2015.02.028

techniques can be utilized to build an IT ontology-based document mining system that collects experience data to share and reuse experience. Experience data is extracted from Frequent Asked Questions (FAQs), public forums, etc. available on the internet and from the IT support team. Collecting, sharing and reusing IT knowledge and problem resolution history helps to detect the root causes of IT problems in a more effective way and brings down the associated cost.

CBR systems are based on the hypothesis that similar problems have similar solutions. CBR technology can be used to develop support systems for business management[1]. Case based reasoning is a computational intelligence tool that solves new problems by using or revising solutions used to solve old similar problems[2]. The new problems need not be solved from scratch, henceforth. An important feature of CBR is its approach towards incremental learning. It learns from experience (past cases). The quality of CBR systems can be improved with the help of data mining technology for feature selection and case retrieval[3]. Data mining, also known as knowledge discovery in databases (KDD), helps in analyzing huge amounts of data and turns it into knowledge[4]. Researchers are focusing on integrating data mining techniques to CBR systems. There is a huge amount of information available to us but the methods of retrieval are still ineffective. Hence, we need to focus more on semantic retrieval to extract meaningful information from documents, thereby increasing precision and recall of retrieval systems. In this paper, the goal is to build an efficient document mining system that gives support for IT team members to provide IT support service that minimizes the interruptions in the smooth functioning of an organization. We propose an IT ontology-based retrieval system that retrieves relevant information from the experience knowledge base.

2. Related works

DUMBO[5] is a system that has applied case-based reasoning to trouble ticket system to diagnose computer network problems. The disadvantage of this system is that the production rules for case matching are manually generated. The development of DUMBO system was influenced by two other systems - MASTER and CRITTER that apply fault management to computer networks and associates CBR in trouble ticket systems. NetPal[6] is a web based dynamic knowledge base system that supports network administrators to troubleshoot tasks, recall and store experience resulting in reduction of time and resources required. Koichi et al. proposed a help desk system that supports the help desk operators respond to enquiries and automate construction of FAQ[7]. In its FAQ building part, a clustering method is proposed to group similar enquiries. Can Bozdogan et al. proposed a system that extracts information from public resources automatically to generate a knowledge base that supports IT management teams by utilizing data mining techniques and information retrieval[8]. The CES+ clustering algorithm used in this system groups and reorganizes the experience data crawled from websites into similar problem/solution pairs. The algorithm is a modified version of that proposed by Koichi et al. and reduces its computational cost. Can Bozdogan et al. later modified the CES+ clustering algorithm with an optimization technique called Multi-objective Genetic Algorithm (MOGA) to automate the process of choosing cluster parameters[9].

The detailed study conducted on data mining and case-based reasoning for IT support service shows that key word-based retrieval has many limitations compared to semantic retrieval. Hence, we need to focus more on approaches for semantic reasoning. In traditional retrieval systems, keywords are used to index documents and queries. The retrieval depends on the keyword overlap of the documents with the query. This can lead to inaccurate results especially when the relevant documents do not share the keywords in the query and the irrelevant documents contain the words in the query. The solution for this problem is semantic retrieval system that understands the meaning or concepts in the users query for efficient retrieval. Ontology defines concepts and their relationships. Tom Gruber defines ‘ontology’ as an explicit specification of a conceptualization that specifies the concepts, their relationships and other definitions that are relevant for modeling a domain[11]. Ontology-based similarity measures quantify the similarity between two concepts in the ontology. The semantic similarity measures can be classified into the following categories[10].

- Path length based measures in which the semantic similarity between concepts is measured by considering the path length and depth of nodes. Examples of path based measures are Rada measure, Bulskov measure, Hirst and St-Onge measure, Wu and Palmer measure and Leacock and Chodorow measure. Depth is considered since similarity is directly proportional to depth. Depth is calculated from root to the target concept.

- Information content based measures which compute similarity between concepts by measuring the information content the concepts share. Examples of information content based measures are Resnik measure, Lin measure and Jiang and Conrath measure.

- Feature based measures in which the properties of concepts are considered to measure the similarity between concepts. Examples of feature based measures are Tversky measure and Pirró measure.

- Hybrid measures which combine knowledge obtained from multiple information sources. The quality of the semantic similarity is improved in hybrid measures since there are alternate information sources if the knowledge derived from an information source is inadequate. Example of hybrid measure is Li measure.

This paper describes the implementation of a retrieval system that identifies the concepts in the FAQ and query by using the IT ontology built using protégé to enable semantic retrieval of documents that are relevant to a given query.

3. Methodology

The first and foremost thing that an IT support team requires is experience to reduce the troubleshooting time and solve problems faster. We explain here an IT experience management system to collect, share and reuse experience so that even less experienced staff and companies with few IT resources can solve problems easily. The system utilizes case-based reasoning, semantic processing and information retrieval techniques, so that previous problem cases can be used to solve problems faster and to share the knowledge and experience of the IT team.

The experience data can be collected from publicly available data on the internet like FAQs, public forums, etc. Web crawlers can be used to obtain such experience data. An experience knowledge base is constructed with the experience data obtained automatically using a crawler and those manually entered by the IT team members. The experience data is structured as problem/solution pairs. To enable semantic retrieval from the knowledge base, we have built an IT ontology using protégé that contains the technical terms related to the information technology domain. This ontology is used to index the FAQ documents and convert query to its concept representation. The concept indexing technique is adapted from the method proposed by Simona et al.[12]. One advantage of the concept indexing technique used in this paper is that the similarity score takes into account the depth of the concepts. Hence, it also considers the fact that the specialized concepts deeper in the taxonomy are more closely related than the higher abstract concepts. The system architecture of the IT ontology-based retrieval system is shown in Fig. 1.

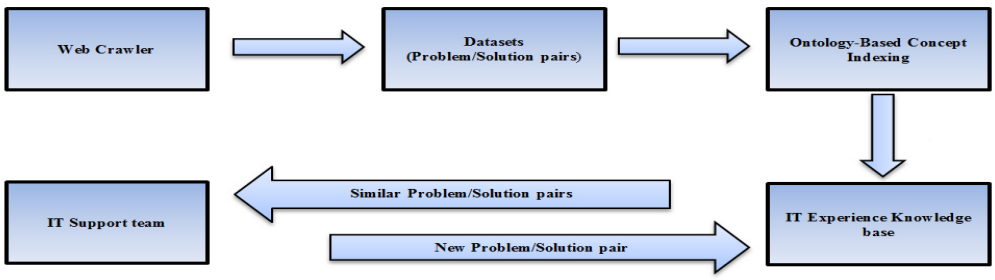

Fig. 1. System Architecture

3.1. Building the IT Ontology

Knowledge resources like ontology are extensively used for word sense disambiguation, natural language processing and informational retrieval. Our system for IT support service is based on the domain specific ontology which includes information technology terms and their relations. IT Ontology is built using protégé 4.3. Protégé is a free, open-source platform to build domain models and knowledge-based application with ontologies[13].

3.2. IT Ontology-Based Indexing

Initially, each document (problem/solution pair) is represented as a collection of the keywords (k1, k2, k3…..kn) that matches the concepts in the IT ontology. All possible combinations of groups of two keywords (km, kx) need to be considered for each document, where x = (m+1) to n and n is the total number of the keywords in the document. For each pair of these keywords, find the lowest common ancestor set (lca set) that consists of three lowest common ancestors (lca1, lca2, lca3) and their similarity scores. The lowest common ancestor (lca) of two keywords is the concept in the IT ontology that has the shortest distance from the two keywords. The similarity score is determined by Wu and Palmers similarity measure as given in equation 1.

$$WPSimScore = 2 * depth(lca) / (depth(kx) + depth(ky)) \quad (1)$$

depth(lca) is the depth of the lowest common ancestor,
depth(kx) is the depth of the keyword kx and
depth(ky) is the depth of the keyword ky.

The common lowest common ancestor list (clca list) for a document is the list containing lca's that are common to more than one lca set.

Example: Suppose a document contains four keywords (k1, k2, k3, k4) that matches the IT ontology terms. Consider all possible combinations of groups of two keywords and then determine the lca set and clca list as shown in Table 1.

Table 1. clca list for a document with four keywords

Keyword pairs	Lca1	Lca2	Lca3	Lca set
(k1, k2)	A1	A2	A3	[A1, A2, A3]
(k1, k3)	A2	A3	A4	[A2, A3, A4]
(k1, k4)	A3	A4	A5	[A3, A4, A5]
(k2, k3)	A2	A3	A4	[A2, A3, A4]
(k2, k4)	A3	A4	A5	[A3, A4, A5]
(k3, k4)	A3	A4	A5	[A3, A4, A5]
clca list = [A2, A3, A4, A5]				

Each pair of keywords is substituted by its first lca that appears in the clca list as shown in Table 2.

The score of a concept is the sum of the Wu and Palmer similarity score of all instances of the concept ci in the document.

$$ConceptScore(ci) = \sum WPSimScore(ci) / n \quad (2)$$

n is the total number of pairs of keywords in the document.

Each document is represented by the concepts and their scores determined using equation 2.

Table 2. Concept determination for a document

Keyword pairs	Lca1	Lca2	Lca3	Lca set	Concept
(k1, k2)	A1	A2	A3	[A1, A2, A3]	A2
(k1, k3)	A2	A3	A4	[A2, A3, A4]	A2
(k1, k4)	A3	A4	A5	[A3, A4, A5]	A3
(k2, k3)	A2	A3	A4	[A2, A3, A4]	A2
(k2, k4)	A3	A4	A5	[A3, A4, A5]	A3
(k3, k4)	A3	A4	A5	[A3, A4, A5]	A3
clca list = [A2, A3, A4, A5]					

3.3. Concept-Based Information retrieval

The ontology-based indexing explained in section 3.2 is used to index the documents (problem/solution pairs) in the experience knowledge base and convert the query entered by the IT team member to its concept representation. Cosine similarity measure determines the similarity between the query and the FAQ documents. This retrieves the relevant documents based on concept similarity. The concept similarity score between the query and each problem/solution pair in the FAQ is computed using equation 3.

$$ConceptSimilarityScore = \alpha * Cossim(query_{con}, problem_{con}) + (1-\alpha) * Cossim(query_{con}, solution_{con}) \quad (3)$$

$$Cossim(query_{con}, problem_{con}) = \frac{\overrightarrow{query_{con}} \bullet \overrightarrow{problem_{con}}}{\|query_{con}\| \|problem_{con}\|} \quad (4)$$

$$Cossim(query_{con}, solution_{con}) = \frac{\overrightarrow{query_{con}} \bullet \overrightarrow{solution_{con}}}{\|query_{con}\| \|solution_{con}\|} \quad (5)$$

$Cossim(query_{con}, problem_{con})$ is the cosine similarity between concept representation of query and concept representation of FAQ problem,

$Cossim(query_{con}, solution_{con})$ is the cosine similarity between concept representation of query and concept representation of FAQ solution and

$0 \le \alpha \le 1$ is used to specify the rate of importance of the content value in problem and solution.

To improve the retrieval performance, a combination of concept similarity score and keyword/term frequency - inverse document frequency (tf-idf) similarity score is considered as shown in equation 6.

$$TotalScore = ConceptSimilarityScore + tfidfSimilarityScore \quad (6)$$

$$tfidfSimilarityScore = \alpha * Cossim(query_{tfidf}, problem_{tfidf}) + (1-\alpha) * Cossim(query_{tfidf}, solution_{tfidf}) \quad (7)$$

$$Cossim(query_{tfidf}, problem_{tfidf}) = \frac{\overrightarrow{query_{tfidf}} \bullet \overrightarrow{problem_{tfidf}}}{\left\|query_{tfidf}\right\| \left\|problem_{tfidf}\right\|} \tag{8}$$

$$Cossim(query_{tfidf}, solution_{tfidf}) = \frac{\overrightarrow{query_{tfidf}} \bullet \overrightarrow{solution_{tfidf}}}{\left\|query_{tfidf}\right\| \left\|solution_{tfidf}\right\|} \tag{9}$$

$Cossim(query_{tfidf}, problem_{tfidf})$ is the cosine similarity between tf-idf weighted vector representation of query and tf-idf weighted vector representation of FAQ problem and

$Cossim(query_{tfidf}, solution_{tfidf})$ is the cosine similarity between tf-idf weighted vector representation of query and tf-idf weighted vector representation of FAQ solution.

Problem/solution pairs that have a *TotalScore* (computed using equation 6) greater than a user specified threshold (γ) are retrieved from the experience knowledge base in the descending order of *TotalScore* values obtained between the query and the FAQ documents.

4. Results

The document mining system for IT support service is implemented using python 2.7.3 on Ubuntu 12.04 LTS. The dataset is an FAQ containing IT problem/solution pairs. The retrieval performances of the document mining system for IT support service using the proposed retrieval approach and the keyword-based retrieval approach have been analyzed and compared. The main problem with the keyword-based retrieval system for IT support service is that documents are not retrieved when there is a mismatch between the vocabulary used by users and that used in the documents. For example, with keyword-based retrieval systems, even though 'sql' and 'structured query language' are equivalent, a query that contains 'sql' will only retrieve documents that have the keyword 'sql'. The documents that contain 'structured query language' and do not have the keyword ‘sql’ are not retrieved. Hence, keyword-based retrieval systems do not take into account the semantic relationship between keywords and meaning of the words/phrases. Precision is the number of relevant documents that are retrieved divided by the total number of retrieved documents. Recall is the number of relevant documents that are retrieved divided by the total number of relevant documents. Precision usually deteriorates as recall increases. The proposed retrieval system that utilizes IT ontology for concept-based indexing and retrieval, solve these problems and thereby, increase precision and recall of the system. The main emphasis is to retrieve meaningfully similar problem/solution pairs from the IT Experience knowledge base. The proposed system retrieves based on the concepts mapped for the query using the IT ontology built and not on the keyword overlap.

Different measures have been proposed for the evaluation of retrieval systems. Mean average precision and average recall-precision curve are the most commonly used measures to evaluate the retrieved results. To analyze the results received in the proposed system, mean average precision (MAP) and average recall-precision curve are used as the performance metrics. Average Precision (AP) is the average of the values of precision acquired for the top k most relevant retrieved documents, at the points at which each relevant document is retrieved. A precision of 0 is assigned for relevant documents that are not retrieved by the system. Mean Average Precision (MAP) is the average of the average precision value for a number of queries. It is a single number that summarizes the retrieval performance of the system and takes into account both the precision and recall of the system[14]. In order to plot the average recall-precision curve, a recall/precision pair is calculated for every point in the ranked list that contains a relevant document. A precision value is then interpolated for each standard recall level (0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0). The rule used to interpolate precision at standard recall level i is to use the maximum precision obtained for any actual recall level greater than or equal to i[15]. Then, average precision is computed at each

standard recall level across a number of queries. The average recall-precision curve plotted helps to evaluate the average performance of the system.

The performance comparison of the two approaches indicate that the proposed approach has high average precision and performs better than the traditional keyword-based approach especially in cases where the keywords of the relevant documents do not match the keywords in the query. Fig. 2(a) shows the average precision values for the concept-based retrieval and the keyword-based retrieval of relevant (related to the topic/concept of the problem) documents for ten queries that share very few keywords with the relevant problem/solution pairs in the experience knowledge base and Fig. 2(b) shows the mean average precision for both approaches. The average recall-precision curves of the IT ontology-based system and the keyword-based system are also plotted for the same set of queries as shown in Fig. 3. It shows the performance improvement of the proposed approach over the keyword-based retrieval approach. The keyword-based retrieval system has good retrieval performance when the relevant documents contain the keywords in the query and the irrelevant documents do not contain the keywords in the query. A vital feature of the proposed retrieval system based on IT ontology is that it retrieves documents belonging to the query topic. For example, a problem related to ubuntu can retrieve not only the ubuntu problem/solution pairs but also the problem/solution pairs on knoppix since both are debian-based linux operating systems. This feature is especially beneficial for an IT support team for problem solving since they can retrieve and compare all the old cases related to the problem topic. Another important feature of the system is that the dimensions of the document vectors are reduced. This reduces the computational cost and inefficiency in document representation.

a

Query	IT Ontology-based approach	Keyword-based approach
Query 1	1.00	0.00
Query 2	1.00	0.00
Query 3	1.00	0.33
Query 4	0.88	0.60
Query 5	0.70	0.05
Query 6	0.60	0.30
Query 7	1.00	0.00
Query 8	0.43	0.00
Query 9	1.00	0.00
Query 10	1.00	0.00

b

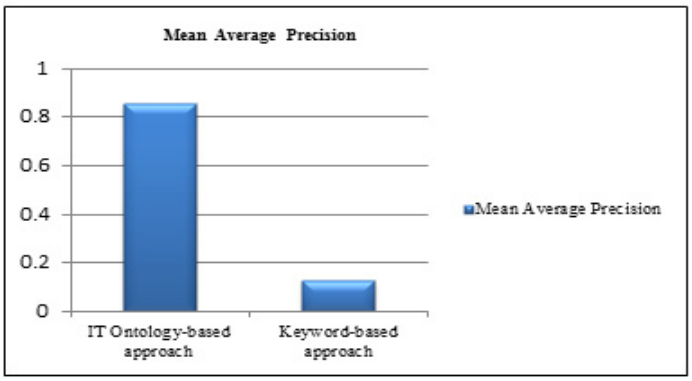

Fig. 2. (a) Average Precision; (b) Mean Average Precision.

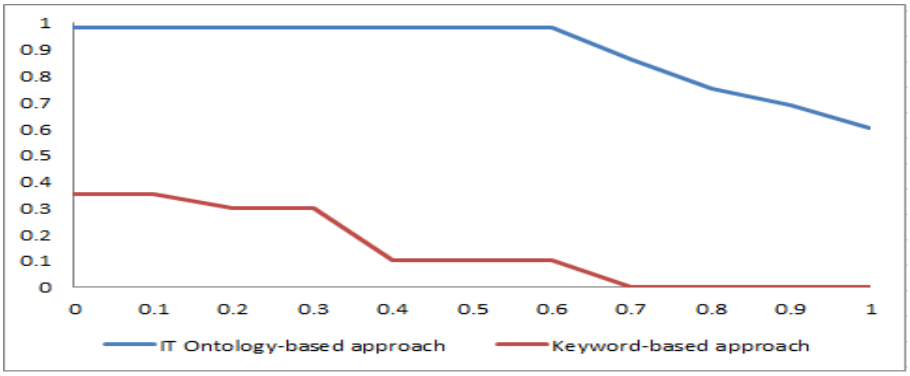

Fig. 3. Average Recall-Precision curve

5. Conclusions and Future Work

In this paper, we have proposed IT ontology-based retrieval system that reduces the burden on IT executives to solve problems and reduces their trouble shooting time. There is a significant difference in the mean average precision (MAP) value and average recall-precision curve of the proposed approach compared to the keyword-based approach for document mining system for IT support service in cases where there is a mismatch between the keywords of the query and the keywords of the relevant documents. This indicates that the retrieval performance of the proposed system is much better than the keyword-based system. The main advantage of the proposed system is that it meaningfully retrieves problem/solution pairs related to the concepts in the query and allows the IT staff to use/adapt old problem cases in problem resolution. Hence, the proposed system that utilizes case-based reasoning, semantic processing and information retrieval techniques has potential benefits over keyword-based access. Future work will focus on the automatic correction of misspelled words in the documents that are similar to words in the IT ontology and automatic IT ontology construction.

References

1. Miguel A. Carmona, Julio Barbancho, Diego F. Larios, Carlos León. Applying case based reasoning for prioritizing areas of business management. *Expert Systems with Applications* 2013;40:3450–3458.
2. Hamid R. Berenji. Case-Based Reasoning for Fault Diagnosis and Prognosis. *IEEE International Conference on Fuzzy Systems* 2006;1318–1321.
3. T. Olsson, P. Funk. Case-based reasoning combined with statistics for diagnostics and prognosis. 25th International Congress on Condition Monitoring and Diagnostic Engineering; *Journal of Physics 2012*.
4. Jiawei Han, Micheline Kamber. *Data Mining: Concepts and Techniques*. 3rd ed. USA: Morgan Kaufmann Publishers; 2006.
5. C. Melchiors, L. M. R. Tarouco. Troubleshooting Network Faults Using Past Experience. *Network Operations and Management Symposium* 2000;549–562.
6. Ashley George, Adetokunbo Makanju, Evangelos Milios, Nur Zincir-Heywood, Markus Latzel, Sotirios Stergiopoulos. NetPal: A Dynamic Network Administration Knowledge Base. *CASCON '08 Proceedings of the 2008 conference of the center for advanced studies on collaborative research.*
7. Koichi Iwai, Kaoru Iida, Masanori Akiyoshi, Norihisa Komoda. A Help Desk Support System with Filtering and Reusing E-mails. *8th IEEE International Conference on Industrial Informatics (INDIN)* 2010;321–325.
8. Can Bozdogan, Nur Zincir-Heywood. Data Mining for Supporting IT Management. *IEEE Network Operations and Management Symposium* 2012;1378–1385.
9. Can Bozdogan, Nur Zincir-Heywood, Yasemin Gokcen. Automatic Optimization for a Clustering Based Approach to Support IT Management. *IFIP/IEEE International Symposium on Integrated Network Management* 2013;1233–1236.
10. Lingling Meng, Runqing Huang, Junzhong Gu. A Review of Semantic Similarity Measures in WordNet. *International Journal of Hybrid Information Technology* 2013.
11. Gruber T. Ontology. *Encyclopedia of Database Systems*. Springer-Verlag; 2009.
12. Simona Barresi, Samia Nefti-Meziani, Yacine Rezgui. A New Conceptual Approach to Document Indexing. *Mexican International Conference on Computer Science* 2009;360–366.
13. http://protege.stanford.edu/
14. Gary Geunbae Lee, Dawei Song, Chin-Yew Lin, Akiko Aizawa, Kazuko Kuriyama, Masaharu Yoshioka, Tetsuya Sakai. *Information Retrieval Technology*. Germany: Springer; 2009.
15. http://trec.nist.gov/