A New Method for Measuring Text Similarity in Learning Management Systems Using WordNet

Bassel Alkhatib, Faculty of Information Technology Engineering, Damascus University, Damascus, Syria & Syrian Virtual University, Damascus, Syria

Ammar Alnahhas, Faculty of Information Technology Engineering, Damascus University, Damascus, Syria & Syrian Virtual University, Damascus, Syria

Firas Albadawi, Faculty of Information Technology Engineering, Damascus University, Damascus, Syria & Syrian Virtual University, Damascus, Syria

ABSTRACT

As text sources are getting broader, measuring text similarity is becoming more compelling. Automatic text classification, search engines and auto answering systems are samples of applications that rely on text similarity. Learning management systems (LMS) are becoming more important since electronic media is getting more publicly available. As LMS continuously needs content enrichment and the web is getting richer, automatic collection of learning materials becomes an innovative idea. Intelligent agents can be used with a similarity measurement method to implement the automatic collection process. This paper presents a new method for measuring text similarity using the well-known WordNet Ontology. The proposed method assumes that a text is similar to another if it represents a more specific semantic. This is more suitable for LMS content enrichment as learning content can usually be expanded by a more specific one. This paper shows how the hierarchy of WordNet can be taken advantage of to determine the importance of a word. It is also shown how similarity method within an e-learning system is exploited to achieve two goals. The first one is the enrichment of the e-learning content, and the second is the detection of semantically similar questions in e-learning questions banks.

Keywords: Intelligent Agent, Learning Management Systems, Semantic Similarity, Text Similarity, WordNet

INTRODUCTION

The web is getting broader by the day and richer contents are getting more available. However, browsing the entire web to collect all useful content is an intractable mission for human beings. So, automatic text similarity bots can be used to search the web for relevant documents.

Much research works have been carried out in this field in the last two decades; some of them employ statisti-

DOI: 10.4018/ijwltt.2014040101

Copyright © 2014, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.

cal methods which are based on pre-classified terms extracted from a corpus. Others depend on semantics and use the natural language processing techniques.

We present in this paper, a new semantic method for text similarity measurement based on WordNet Ontology and suitable for learning management systems. Section II shows a brief summary of some previous related works. Sections III and IV describe the proposed method and its algorithm. Section V shows how the method is applied to enrich LMS content. Section VI describes how to use the similarity method to detect similar questions in questions banks. The results are reviewed in section VII and the paper is finally concluded with section VIII.

LITERATURE REVIEW

Many methods have been presented to measure text similarity. Traditional methods are based on text lexical analysis and adopted by many information retrieval systems to find similar texts based on a text query. Some new research works are based on corpora-extracted statistics, and are considered to be statistically oriented (Mihalcea, Corley, & Strapparava, 2006; Corley & Mihalcea, 2005; Islam & Inkpen, 2008; Amala Bai & Manimegalai, 2013). Many other studies have focused on the concepts of texts, where some conceptual representations like ontologies is used to determine the overwhelming concepts of a text (Pandya & Bhattacharyya, 2005; Wang & Taylor, 2007). Some other works are based on machine learning techniques, where an agent is used to learn how to test text similarity (Bilenko & Mooney, 2003; Lee, Pincombe, & Welsh, 2005). Some hybrid systems are also proposed such as the one in Mohle and Mihalcea (2009).

Using Ontologies in e-learning systems were presented in many researches, as in Henze, Dolog, and Nejdl (2004) where the authors proposed a method to personalize e-learning contents using Ontologies and semantic web resources. They investigate a logic-based approach to educational hypermedia using TRIPLE, which is a rule and query language for the semantic web.

Many other researchers used WordNet in e-learning, Carbonaro (2010) proposed a research that aims to build a summarization system to support tutors in managing student communication and interaction within an educational environment. They show that Concept-based approaches to represent dynamic and unstructured information can be useful to address issues such as trying to determine the key concepts and to summarize the information exchanged within a personalized environment. It seems a promising technology for implementing a distance learning environment; enabling the organization to deliver learning materials around small pieces of semantically enriched resources.

The study in Hung and Yee (2005) shows a semantic-based automated question answering system that can act like a virtual tutor to answer student questions online. This system, not only relieves the tutor from the burden of answering many questions, but also allows students to get answers promptly without waiting for the tutor‘s response.

Another research example of using Ontology in e-learning is Deline, Lin, Wen, and Gašević (2009). This research proposed an ontology-driven software development methodology which is appropriate for intelligent ontology-driven

Copyright © 2014, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.

systems. This approach employs ontologies as key execution components such as e-Advisor, and is biased towards an integration of incremental and iterative ontology development. The researchers concluded that the benefits of ontologies for intelligent educational systems are that intelligent agents can use the developed ontologies as the basis for their knowledge base construction, reasoning and interface design. They used e-Advisor development as a case study, specifically how ontologies are developed and maintained. In the e-Advisor architecture, the ontologies formally define domain entities and the relations among them. Based on the ontologies, the structural part of the knowledge base is modeled.

Compared to previous researches, our proposed work will employ the concept extraction method using WordNet Ontology to find semantically similar documents representing the same learning content.

PROPOSED APPROACH FOR TEXT SIMILARITY MEASUREMENT

WordNet® is a large lexical database of English nouns, verbs, adjectives and adverbs; each is grouped into sets of cognitive synonyms (synsets). Each synset expresses a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations (Miller, 1995).

An English word may have more than one synset in WordNet, each for every concept it represents. We mark these synsets as "containing the word", hence for each synset S containing the word W we note:

$$con\left(S,W\right)$$

Where con is a relation from the set of all synsets of WordNet to the set of all English vocabulary.

Each synset in WordNet, except for the root (entity), is related to another synset with a hyponym relation (i.e. the parent). If we denote this relation as hyp, we find that:

$$\left[hyp\left(S1,S2\right)and\,con\left(S1,W\right)\right]\rightarrow con\left(S2,W\right)$$

The hyponym relation is a generalization relation; thus the meaning of the hyponym contains the meaning of its sub-synsets. Therefore, if a synset contains some word, its hyponym should contain the same word (See Figure 1).

Since the goal of this approach is to measure the semantic similarity between two texts T_1 and T_2. Also, as a text is a series of words; consequently, these texts should be tokenized to a list of words.

We could express T_1 and T_2 as follow:

$$T1=\left[W1,W2\ldots Wn\right]$$

$$T2=\left[W1,W2\ldots Wm\right]$$

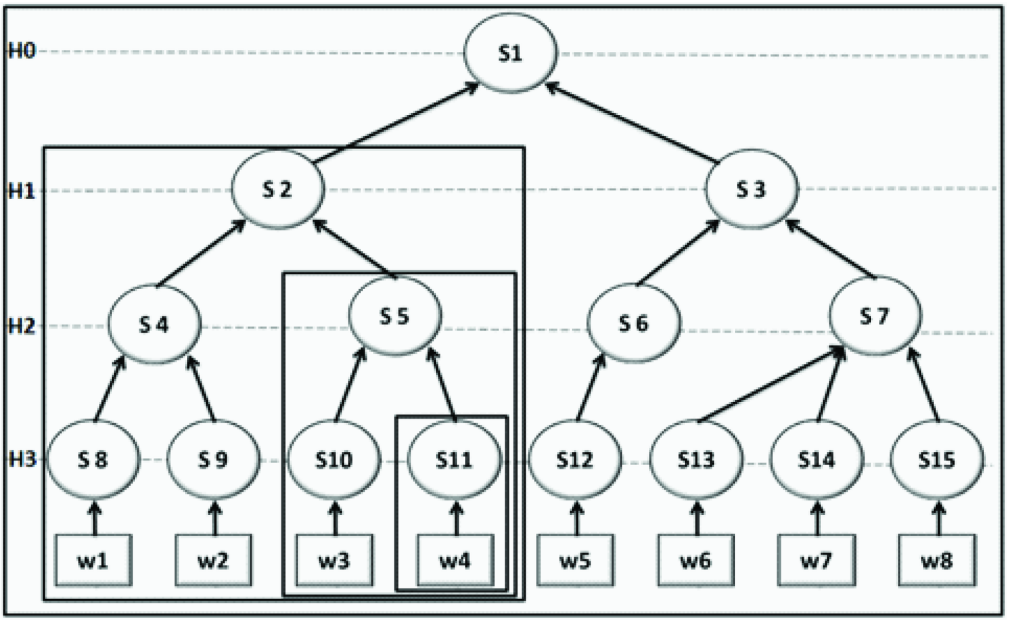
Now we can define the coverage cov(S, T) of a synset S in WordNet to some text T as the number of words this synset contains of that text:

$$cov\left(S,T\right)=count\left(con\left(S,W\right)\right)for\,each\,W\,T$$

It is clear that the larger the coverage is, the more the synset expresses the meaning of the text.

Copyright © 2014, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.

Figure 1. Hyponyms contain the same words as sub-synsets



Nevertheless, it is useless to consider this factor apart from other measures. This stems from the fact that the synset "entity" which is the top synset of WordNet will therefore express all text semantics equally, as it is a hyponym of all synsets in WordNet.

Another important factor to consider is the degree of generalization a synset has. We define the height (H) of a synset as the minimum number of hyponym relations that connects this synset with the root WordNet synset "entity". Clearly, the larger the height, the more specific the synset is, since the hyponym relation is a generalization relation. Therefore, as we go deeper in the net, synsets that are more specific are reached.

This work intends to find a set of synsets that perfectly represents a text with as much precision and recall as high as possible. Therefore, we are looking for synsets that have large coverage with most specific meaning. Larger coverage means that a synset expresses larger part of a text and it also increases the precision of the results. Whereas being more specific helps in distinguishing a text and making it easier to represent apart from another one, this increases the recall of the results.

We define the importance imp(S, T) of a synset S representing a text T as the sum of the coverage cov(S, T) with H(S) representing the degree of generalization:

$$imp\left(S,T\right)=\frac{cov\left(S,T\right)}{N}+p*H\left(S\right) \qquad (1)$$

Where N is the total number of words in T, and p is a normalization parameter that its value is set experimentally. Larger p values make the algorithm tends to increase the importance of more specific words, whereas small values give more importance to the coverage.

Copyright © 2014, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.

Clearly, the more the importance of a synset is, the more the synset suites the text. We define the "Digest" of a text as the set of L highest importance value synsets, i.e. if the descending ordered list of synsets according to their importance value is:

$$set = \{S1, S2, \ldots, Sn\}$$

Where S_1 is the highest importance value synset and S_n is the lowest importance value synset, then the digest dig(T) of text T is a set defined as follow:

$$dig(T) = \{Si\} \; where \; Si \in set \; and \; i \leq L \quad (2)$$

Now each text T is represented by its digest dig(T) which is semantically expressing T (See Figure 2).

To measure the similarity between text T_1 and text T_2 we can just measure the distance between their digests dig(T_1) and dig(T_2) which can be achieved as follows:

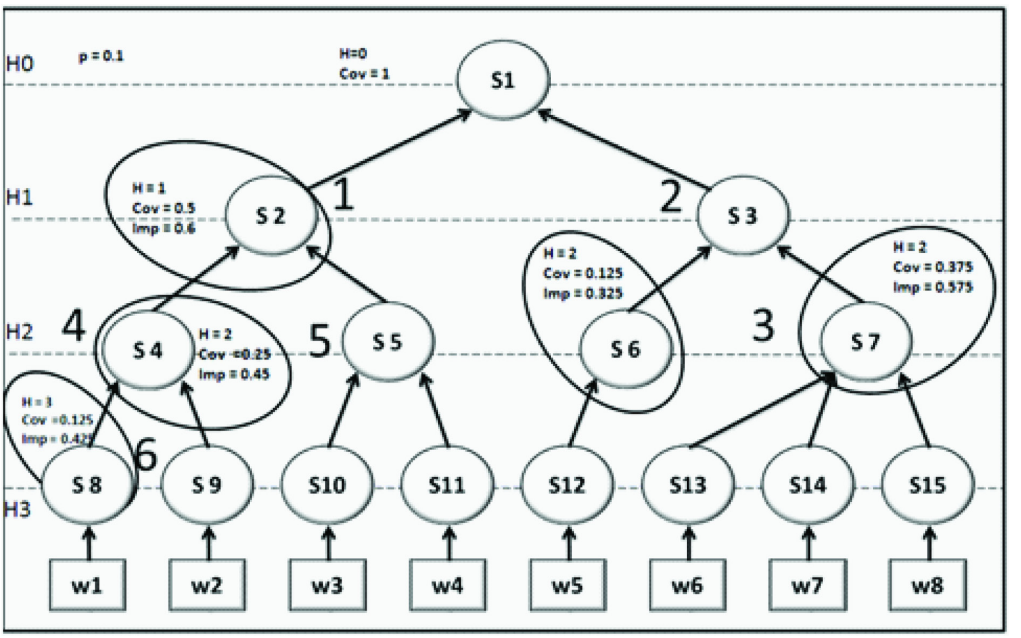
In order to define the distance between two synsets; it is intuitively considered that two synsets are close if one has a more general concept of another. The only semantic relation used in WordNet is the hyponym relation. Hence, if two synsets related with a direct or indirect hyponym relation, then they are considered to be similar.

Formally, the function dis can be defined as:

$$dis : S \times S \rightarrow R$$

Where S is the set of all synsets in WordNet. The following recursive

Figure 2. The digest is the highest ordered important synsets



Copyright © 2014, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.

function dis(S_1,S_2) denotes the distance between two synsets S_1 and S_2:

$$dis\left(S_1,\ S_2\right)=\begin{cases}1,\ S_1=hyp\left(S_2\right)\\ \infty,\ S_2=\text{"}entity\text{"}\\ 1+dis\left(S_1,\ hyp\left(S_2\right)\right),\ else\end{cases}$$

If there is no hyponym route between S_1 and S_2, then the distance is ∞. Otherwise, the distance is the number of hyponym relations between S_1 and S_2.

As the similarity is the inverse of the distance, the similarity between two synsets S_1 and S_2 can be defined as inversely proportional to the distance, if the similarity is denoted as sim(S_1, S2), then:

$$sim\left(S1,S2\right)=max\left(0,1-e*dis\left(S1,S2\right)\right) \quad (4)$$

Where e is a scaling factor with a value in the range[0, 1/d], where d is the maximum accepted distance between two similar synsets (dis(S_1, S_2)). Larger values for e decrease the distance on which synsets are considered similar, whereas smaller values enlarge the same distance.

Now, the similarity between two text digests can be defined as follow:

$$sim\left(D_1,D_2\right)=\frac{1}{2\mathrm{L}}\sum_{S_1\in D_1}\mathrm{Max}_{S_2\in D_2}sim\left(S_1,S_2\right)+\frac{1}{2\mathrm{L}}\sum_{S_1\in D_2}\mathrm{Max}_{S_2\in D_1}sim\left(S_1,S_2\right) \quad (5)$$

Consider as an example the sample shown in Figure 3 and Table 1 where distances are (e=0.5).

$$sim\left(D1,D2\right)=\frac{1+1+1+0.5}{8}+\frac{0.5+1+1+1}{8}=0.875$$

THE TEXT SIMILARITY MEASUREMENT ALGORITHM

This section, introduces the practical algorithm employed in text similarity measurement. The two texts to be compared should be tokenized first; and stop words are excluded from the token set. Then each word is processed separately by looking up all synsets this word belongs to in WordNet Ontology. Then each of them is used to extract the series of all hyponym synsets starting from the in-process synset to the root of WordNet. Each resulting synset is assumed to be a potential digest member therefore; its coverage increases, as it represents the token it was generated from. This process continues until all tokens are processed.

The main processing algorithm is as follows in Box 1.

To process each synset, the synset is checked to verify its existence in the list of previously processed synsets. The synset coverage is updated; otherwise, the synset is added (Box 2).

The algorithm is clearly recursive and should terminate when reaching the root of WordNet.

The height of each synset is calculated recursively as well; finally, the digest is chosen to be the synsets with the highest importance values.

Copyright © 2014, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.

Figure 3. Finding similarity between two texts

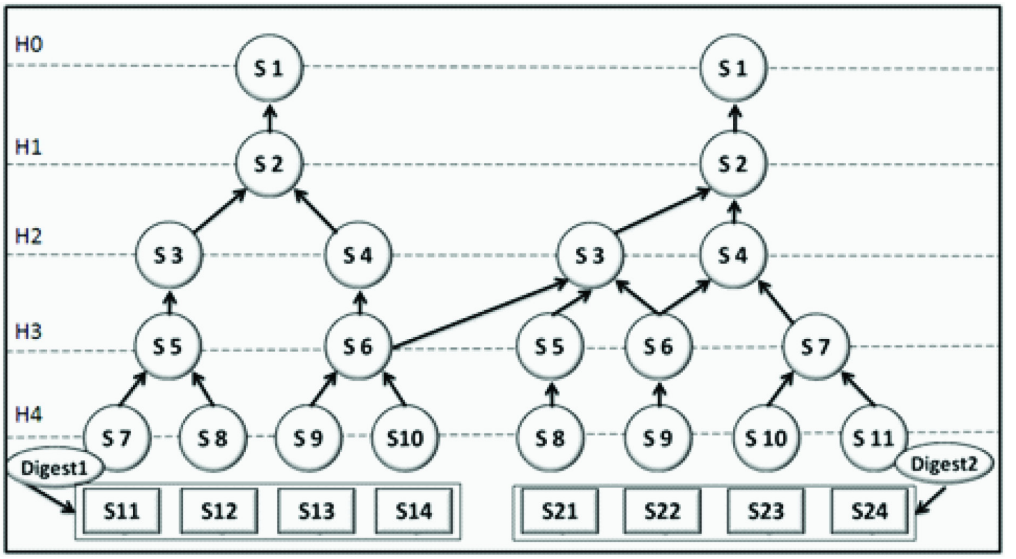


Table 1. e=0.5 distance

	S7	S8	S9	S10
S8	0	1 – 0.5 * 0 = 1	0	0
S9	0	0	1 – 0.5 * 0 = 1	0
S10	1 – 0.5 * 1 = 0.5	0	0	1 – 0.5 * 0 = 1
S11	1 – 0.5 * 1 = 0.5	0	0	0

Box 1.

```
SynsetSet = {}
For each word W in set of tokens
     For each synset S in synsets of W
          Set con(S, W)
          SynsetSet = SynsetSet + S
          Process(S, SynsetSet)
```

Box 2.

```
Process(S, SynsetSet)
     If SynsetSet contains S then
          cov(S, T) = cov(S, T) + 1
     Else
          cov(S, T) = 1
     Let SH be the hyponym of S
            Process (SH, SynsetSet)
```

Copyright © 2014, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.

Using Text Similarity to Enrich LMS Content

LMS initially has some content that is added by tutors. The system will use an intelligent agent to search the web for more academic content. This content is semantically related to already existing content which is listed in the educational syllabus. The criterion specifying whether the content is useful depends on the text similarity measurement proposed in this paper.

Practically, a web crawler is used to navigate through a set of selected educational websites. If a webpage is acquired with an article inside, the text is stripped out of the HTML. Then the stripped text is checked against all existing articles in the LMS academic content. The similarity between the candidate text and each existing article is measured, and the average of these similarities is calculated. Formally, if a new candidate text is denoted as T, the texts of the existing articles as ET_i and the number of these articles as n, then TS is calculated as follows:

$$TS = average\left(sim\left(T, ETi\right)\right) for\ all\ i <= n \quad (6)$$

A threshold TH is chosen to be the minimum value of similarity for the new text to be accepted. Hence if TS is greater than TH, then the text is accepted, otherwise it is rejected.

Figure 4, shows the main modules of a proposed system to test our proposed method. The crawler sends the HTML content it finds in a web page to an HTML parser. The parser extracts the text then it sends it to another component to find the useful text only. The Readability project (www.readability.com) is used since a text may have many useless contents such as ads, slogans or other materials. The text is sent to the main processing unit, which is the digest extractor unit. This unit uses the ontology manager to handle the WordNet files. The resulting digest is compared to the digest of each existing article stored in the LMS's academic database. Finally, if the candidate text meets the acceptance condition proposed, it is stored in the academic database.

Figure 5, shows the integration of the proposed system within e-learning process. Tutor has to enter some topics or documents and associate websites links for the system to crawl on. The developed similarity module will then surf all the provided links and apply similarity measurement in order to get similar documents. Tutor has to check the documents' relevance to the desired needs before integrating the retrieved document within the educational content.

Figures 6 and 7 show screenshots of how a tutor can enter some text and validate results.

USING TEXT SIMILARITY TO DETECT SIMILAR QUESTIONS

The similarity method is exploited for the detection of similar questions in e-learning question banks.

This problem arises usually when a group of tutors work on enriching the same questions bank as they add multiple questions covering the same outcomes using different vocabulary. Hence, it becomes necessary to check this kind of semantic redundancy.

Copyright © 2014, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.

Figure 4. System Structure

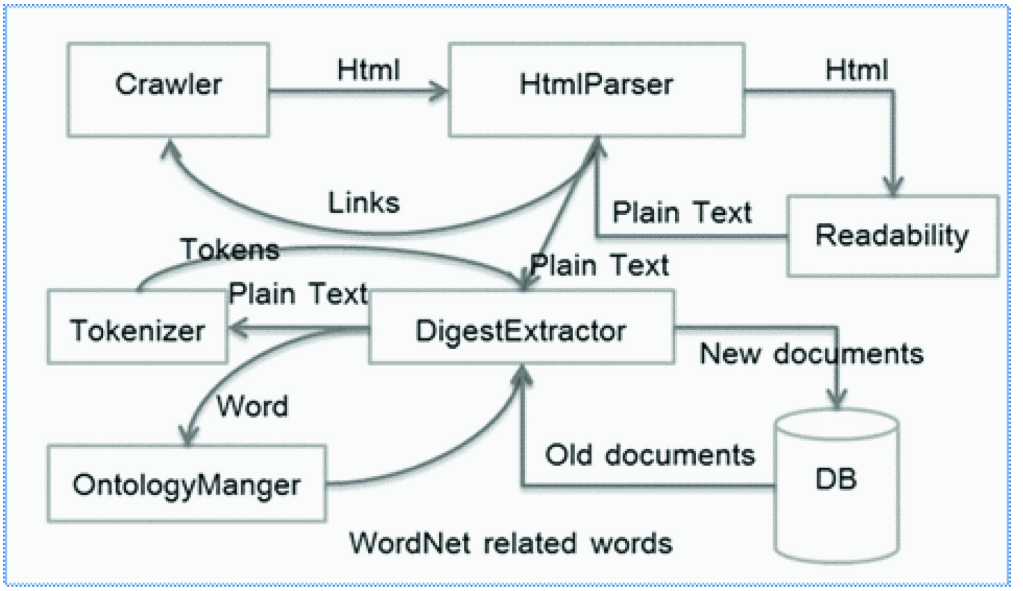
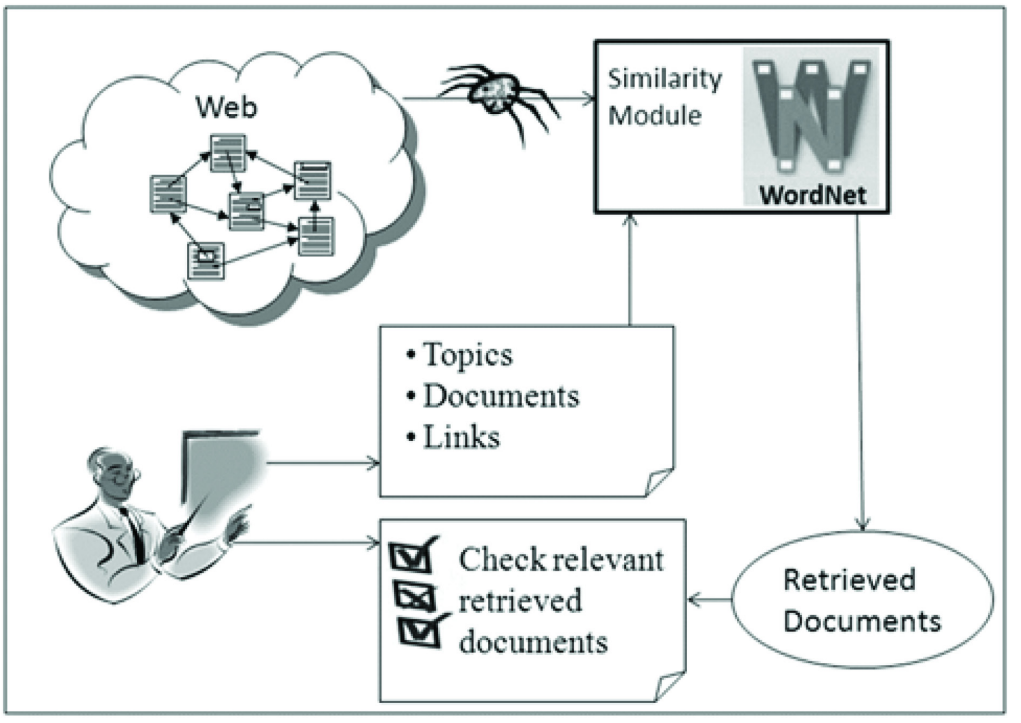


Figure 5. Integration of Similarity module within e-learning process



Copyright © 2014, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.

Figure 6. Tutor can enter some text

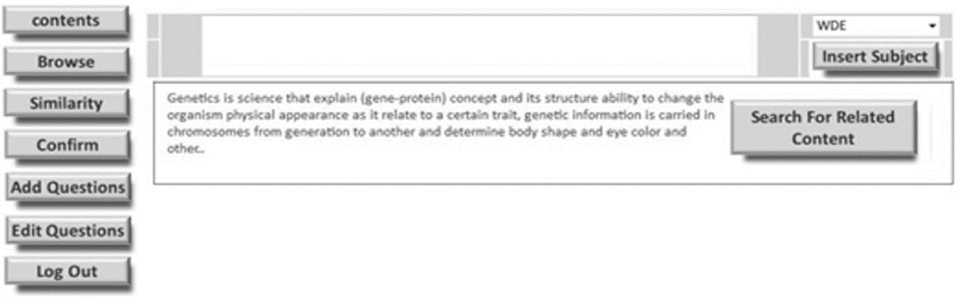
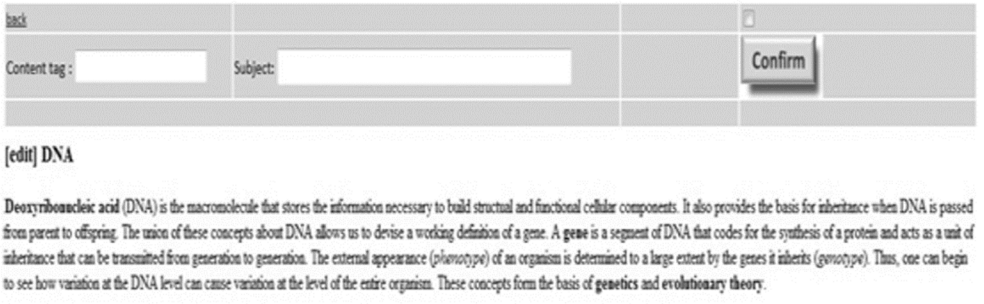


Figure 7. Tutor has to validate retrieved results



Since the proposed method is looking for similarities among texts, it is normal in case of questions to be more adequate, especially that questions usually are relatively short and belong to a specific domain. Therefore, a limited set of words that share the same domain, and thus may lead to a high detection ratio as words of similar meaning may commonly share the same or similar route in WordNet.

Practically, applying the developed method showed good results in detecting most of the similar questions.

TESTING AND RESULTS

To present the results of the proposed method, a practical sample is used to demonstrate the following two sentences:

- I like to play basketball
- Jogging sport is good for health.

The first sentence is tokenized and its stop words are removed, the result is [play, basketball]. The table below shows some of the synsets of the WordNet that are extracted for each token of the first sentence. The coverage, height and the importance of each synset is shown.

Copyright © 2014, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.

Recalling the importance Formula (1) with p=0.6 gives the results in Table 2 and Table 3.

The same steps are applied to the second sentence and its table is shown below:

Setting the digest length L to 6, we take the 6 highest important synsets from both tables to be the digests for the two sentences. Then, the similarity is calculated by applying the Formula (5) in section III, as shown in Table 4. The similarity measurement result is 0.76.

The proposed similarity method is applied to a real e-learning system. Initially, the system had 50 articles in

Table 2. p=0.6

	Coverage	Height	Importance
play	1	7	1/2 + 7 * 0.6 = 4.7
basketball	1	7	4.7
diversion	2	6	4.6
ball	1	6	4.1
action	1	6	4.1
motion	1	6	4.1
court_game	1	6	4.1
attempt	1	6	4.1
basketball_equipment	1	6	4.1
movability	1	6	4.1
. . .			
entity	2	0	1

Table 3. Second changes

	Coverage	Height	Importance
sport	2	6	2/ 4 + 0.6 * 6 = 4.1
jogging	1	6	3.85
good	1	6	3.85
health	1	6	3.85
diversion	2	5	3.5
football_play	1	5	3.25
athlete	1	5	3.25
wellbeing	1	5	3.25
condition	1	5	3.25
operation	1	5	3.25
. . .			
entity	4	0	1

Copyright © 2014, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.

Table 4. Formula applied

	sport	jogging	good	health	diversion	football_play	MAX
play	1	0.9	0	0	0.95	0.95	1
basketball	0.85	0	0	0	0.8	0	0.85
diversion	0.95	0.85	0	0	1	0	1
ball	0.7	0	0	0	0.65	0	0.7
action	0	0.9	0	0	0.75	0	0.9
motion	0	0.9	0	0	0	0	0.9
MAX	1	0.9	0	0	1	0.95	

$$\text{similarity} = \frac{1+0.85+1+0.7+0.9+0.9}{12} + \frac{1+0.9+0+0+1+0.95}{12} = 0.76$$

its academic content (seed documents) that are added by tutors. The similarity module crawled the web and chose 164 academic articles to be added. Tutors approved 140 articles from the automatically retrieved ones, which leads to an accuracy of 85.36% for the proposed method.

CONCLUSION AND FUTURE WORK

This article presented a new method for measuring text similarity. It was shown how this method could be used for many purposes in e-learning systems.

The evaluation of the developed method proves its validity in e-learning systems, nevertheless, depending on single word similarity may have some drawbacks. One of which is the variety of meanings of a single word, as it may increase the ambiguity and needs to disambiguate the difference meanings of the word. However, the attained results have not been much affected by this problem as words have usually specific meanings in a particular domain. Moreover, the tutor's participation in this approach in checking the validity of the results will play an important role in decision making.

The empirical testing of the system has shown interesting and encouraging results, strengthening our work hypotheses. Our future work aims to take into account the semantic structure of sentences to generalize our search domain.

REFERENCES

Amala Bai, V. M., & Manimegalai, D. (2013). A Document Level Measure for Text Categorization. *International Review on Computer and Software. Praise Worthy Prize*, *8*(6), 1374–1381.

Bilenko, M., & Mooney, R. J. (2003). Adaptive Duplicate Detection Using Learnable String Similarity Measures. (pp. 39-48). *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery*. doi:10.1145/956750.956759

Carbonaro, A. (2010). WordNet-based Summarization to Enhance Learning Interaction Tutoring. *Journal of e-Learning and Knowledge Society*, http://je-lks.org/ojs/index.php/Je-LKS_EN/article/view/413

Copyright © 2014, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.

Corley, C., & Mihalcea, R. (2005). Measuring the Semantic Similarity of Texts. (pp. 13-18). Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment. Association for Computational Linguistics.

Deline, G., Lin, F., Wen, D., Gašević, D., & Kinshuk, . (2009). A Case Study of Ontology-Driven Development of Intelligent Educational Systems. *International Journal of Web-Based Learning and Teaching Technologies*, *4*(1), 66–81. doi:10.4018/jwltt.2009010105

Henze, N., Dolog, P., & Nejdl, W. (2004). Reasoning and Ontologies for Personalized E-Learning in the Semantic Web. *Journal of Educational Technology & Society*, *7*(4).

Hung, J., & Yee, G. (2005). Applying Word Sense Disambiguation to Question Answering System for e-Learning. *19th International Conference on Advanced Information Networking and Applications (AINA'05),* 1, pp. 157-162. doi:10.1109/AINA.2005.121

Islam, A., & Inkpen, D. (2008). Semantic Text Similarity Using Corpus-Based. Word Similarity and String Similarity. *ACM Transactions on Knowledge Discovery from Data (TKDD) 2.2, 10*.

Lee, M. D., Pincombe, B., & Welsh, M. B. (2005). An Empirical Evaluation of Models of Text Document Similarity. *XXVII Annual Conference of the Cognitive Science Society* (pp. 1254-1259). Cognitive Science Society.

Mihalcea, R., Corley, C., & Strapparava, C. (2006). Corpus-based and Knowledge-based Measures of Text Semantic Similarity. American Association for Artificial Intelligence, 6.

Miller, G. A. (1995). A Lexical Database for English. *Communications of the ACM*, *38*(11), 39–41. doi:10.1145/219717.219748

Mohle, M., & Mihalcea, R. (2009). Text-to-text Semantic Similarity for Automatic Short Answer Grading. Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics. Association for Computational Linguistics.

Pandya, A., & Bhattacharyya, P. (2005). Text Similarity Measurement Using Concept Representation of Texts. *Pattern Recognition and Machine Intelligence. Springer Berlin Heidelberg*, 678-683.

Wang, J. Z., & Taylor, W. (2007). Concept Forest: A New Ontology-assisted Text Document Similarity Measurement Method. (pp. 395-401). *IEEE/WIC/ACM International Conference on Web Intelligence*. doi:10.1109/WI.2007.11

Bassel Alkhatib is the Web Science Master Director at the Syrian Virtual University and the head of Artificial Intelligence Department at the Faculty of Information Technology Engineering- Damascus University. He holds a PhD degree in computer science from the University of Bordeaux-France, 1993. Dr. Alkhatib supervises many PhD students in Web Mining, and Knowledge Management. He also leads and teaches modules at both BSc. and MSc. levels in computer science and web engineering in both the Syrian Virtual University and Damascus University.

Ammar Alnahhas is a teaching assistant at the Syrian Virtual University and Damascus University. He has a Master degree in Artificial Intelligence from Damascus University 2011. He has achieved many projects in the field of AI and Web Engineering. He teaches fuzzy logic, expert systems and many other subjects in both SVU and Damascus University.

Firas Albadawi is a Web Science master student. He has a Bachelor degree in Information Technology, SVU 2009.

Copyright © 2014, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.