A Comprehension-based Framework for Measuring Semantic Similarity

Naser S Al Madi, Javed I Khan
Department of Computer Science
Kent State University
Kent, USA
nmadi@kent.edu, javed@cs.kent.edu

***Abstract*—We present a comprehension-based framework for measuring semantic similarity between documents of text. In various situations, vector-based similarity measures fail to capture deep semantic relations between terms. Our computational comprehension model processes textual content in a way that resembles human readers, paying attention to context, location, and acquisition time of semantic concepts. The model extracts key semantic structures that are representative of the document. These semantic structures are compared using the WordNet WUP measure giving a Semantic-similarity score of the processed documents. Three experiments are illustrated comparing our results with three popular vector-based similarity measures and human readers. Our framework provided correct results in cases where vector-based methods fail. These results highlight the importance of using computational cognitive methods, such as comprehension models, in semantic analysis and text mining.**

I. Introduction

Semantic similarity measures are vital in modern day technologies. The applications span over document clustering and classification [1], text analysis [2][3], biomedical ontologies [4], classification stem cell research [5], and information retrieval [6]. Standard similarity measures, such as Latent Semantic Analysis (LSA) [7], have served as the main pillar for text similarity for a long period of time. Such methods treat a document as a bag of words, giving no regard to context and order of appearance. In addition, LSA requires a large corpus of text to learn co-occurrences of terms, and that might not be available in a web search query for example [2]. Last, traditional similarity measures treat words as isolated terms neglecting the semantic relations in a document.

In this work, we describe a framework that is able to measure semantic similarity between documents. The proposed framework is scalable and provides good results with small and large documents. In addition, our framework pays special attention to context, location, and acquisition time of each semantic concept in a document. Most importantly, the framework depends on a cognitive model of comprehension that processes documents in a similar way to humans, and constructs a semantic concept-network that represents the document and the relations between the contained concepts. The weighted concept-network representation has the advantage of minimizing redundancy, for example the terms UN and United Nations are represented as one concept. In addition, due to the weights in the concept network, semantic relations between concepts can be analyzed, measured, and mined for insights about the document, such as [8] [9]. The use of a cognitive model of comprehension gives a great advantage to the framework, as it mimics the performance of a human reader [10]. At the same time, the use of Natural Language Processing and a semantic lexicon add to the versatility of the frame work, as it can be used with structured and unstructured documents. We Present three experiments presenting the three possible cases: comparing a document to itself, comparing two documents with the same topic, and comparing two documents with unrelated topics. In addition, we compare our results with three popular vector-based similarity measures, and we compare our method to a group of human readers.

Some of notable related works include [9], which proposed a machine-learning measure of similarity. The advantages in that approach include the use of concepts as a representational model, and giving attention to context and semantic relations between concepts. Similarly.While [11] gives a good review of the advantages of semantic measures of similarity, and uses topic maps as a representational model. In addition, [2] considers the case of small text, and proposes a measure of similarity that is suitable for web search queries and information retrieval. On the problem of clustering unstructured documents, [1] provides a solution that depends purely on the K-means clustering algorithm.

A. Cognitive Theory of Text Comprehension

An educated person has the ability to decode print into meaningful information through the process of comprehension. This unique and acquired skill for humans [12] has been the focus of research for a number of years. One of the influential theories on text comprehension is the Constriction-Integration Model (CI-Model) [13]. The theory suggests that text comprehension is performed through two sequential processes namely construction (sometimes segmentation) and integration. Construction describes reading the text and building a linguistic level mental representation of the text. Integration describes integrating the newly created representation with prior knowledge (long-term memory). That understanding suggests that comprehension is an iterative process of recognizing concepts and gradually creating connections or associations between new concepts and existing ones.

978-1-5090-2508-4/17/$31.00 ©2017 IEEE

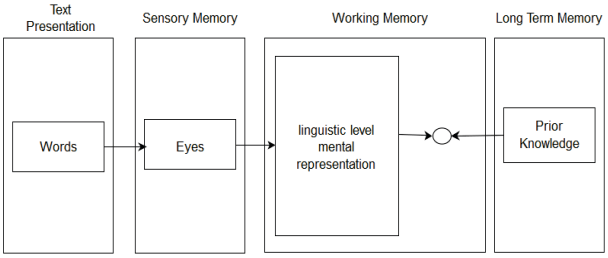

Fig. 1. Model of Text Comprehension

Text comprehension theories often describe the outcome of comprehension in terms of a mental representation. The work in [14] describes perceptual representations, verbal representations, and semantic representations for comprehension. In our study, we focus on the semantic representation of comprehension, therefore the relationship between concepts and associations can be represented as a semantic network. Our work is based on [10] which provides a framework for modeling the growth of semantic networks during comprehension and provide a quantitative measure of concept learning.

II. Methods

In this section, we present the details of our framework. We can review the steps involved in our method briefly, starting with an unstructured document that is processed Natural Language Processing extracting key semantic concepts in each sentence. Next, the comprehension model mimics a human reader and creates a weighted semantic network representation of that document, paying special attention to concept acquisition time. The weights in the semantic network reflect the importance of concepts in the document. Finally, The weighted semantic network is processed by the semantic similarity module where the most important concepts of each document are compared semantically. The output of the last step is a semantic-similarity score between 1 and 0, where 0 represents no semantic-similarity, and 1 representing semantically identical documents.

A. Natural Language Processing

Natural Language Processing (NLP) plays an important role in the process of recognizing concepts in a given text. Therefore, we use Stanford CoreNLP tool [15] to extract key semantic concepts in a document, by removing non-functional terms and stop-words from the text. A sample sentence and the correlating extracted concepts are shown in Figure 2.

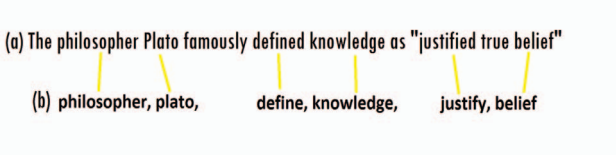

Fig. 2. (a) Sample sentence from Wikipedia. (b) The extracted concepts using CoreNLP.

B. Computational Comprehension Model

Using a framework of concept learning [10], the model takes an unstructured text document and constructs a weighted semantic network representing the relationships between concepts. Additionally, this weighted network can be analyzed turning quantitative network metrics into qualitative insights as demonstrated in [8]. This allows us to construct a semantic network of any text document and analyze it to get the topic, context, and the most important concepts in that document. This could be useful when studying topic evolution, such as the work in [16].

To illustrate how the model works and how the meaning of a text is converted into a semantic concept network, Figure 3 shows a sample concept network of a three sentences text. In this example, each sentence is presented in a single unit of time (episode). Figure 3 corresponds to the following sample text:

> "Knowledge is a familiarity. Awareness or understanding of something. Such as facts."

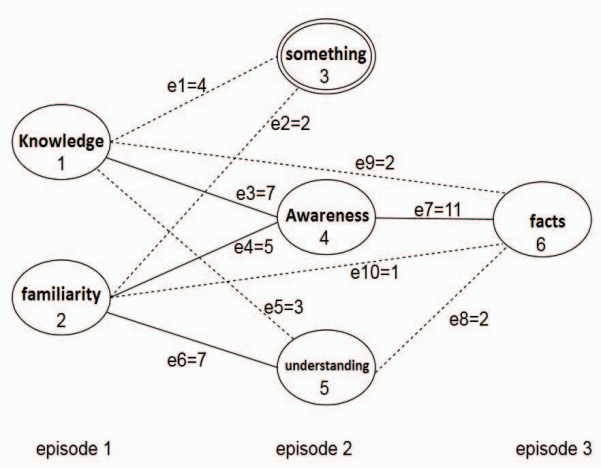

Fig. 3. Sample Concept Network.

In the first episode, the reader recognized the concepts "Knowledge" and "familiarity". In episode two, the reader only recognized "Awareness", "understanding", and missed "something". In episode three, the reader recognized "facts". As for associations, the reader missed the association between "Knowledge" and "understanding". Also, the association between "understanding" and "facts" was missed too.

Concepts are represented as nodes and associations between concepts are represented as edges. The weights in Figure 3 represent association strength between concepts. One important feature of this model is that it pays special attention to concept acquisition time. For that, time frames are represented by episodes, where concepts acquired in episode one are considered background knowledge for concepts in episode two and so on. This means that discovering concepts in an episode depends on what was discovered in all previous episodes. In this model, the ability to recognize a concept is represented in the concept-recognition threshold (S), which determines whether a concept can be discovered or not based

on total associations strength to that concept. For example, in figure 3 assuming that the concept recognition threshold (S) is seven. The total weights of associations pointing to concept three ("something") is six (e1 + e2 < S), which is less than the threshold, and hence the reader was unable to recognize concept three. At the same time, the total weights pointing at concept four ("Awareness") is twelve, which is higher than the threshold, and hence the reader could discover that concept. Similarly, association threshold (I) determines whether an association is created between two concepts based on the weight of that association. For example, in Figure 3 the association threshold is five. Therefore, the association between concept one and concept five was not discovered.

In the previous example the weights and thresholds were provided, but in a semantic network that was constructed from an individual document (ICN) we do not have any weights or thresholds. How do we get these weights and thresholds? ICNs can be represented as a set of inequalities, similar to what we saw in the previous example. The inequalities set an upper bound and a lower bound for concept recognition threshold (S) and association threshold (I). From the example in Figure 3, we can draw the following inequalities:

Recognition Threshold (S):

$$e1 + e2 \leq S$$
$$e3 + e4 \geq S$$
$$e5 + e6 \geq S$$
$$e9 + e10 + e7 + e8 \geq S$$

Association Threshold (I):

$$e1 \leq I; e1 > 0$$
$$e3 \geq I; e3 > 0$$
$$e5 \geq I; e5 > 0$$
$$e2 \leq I; e2 > 0$$
$$e4 \leq I; e4 > 0$$
$$e6 \geq I; e6 > 0$$
$$e7 \geq I; e7 > 0$$
$$e8 \leq I; e8 > 0$$

Linear programming is used to find suitable values for all variables to satisfy the inequalities. The resulting variable vector contains the weights for nodes and associations. The resulting network contains concepts and relationships between concepts, and weights describing the importance of each concept and association.

C. WordNet and Semantic Similarity Measures

The output of the comprehension model is a weighted semantic network containing weights for each concept. The most important concept is the most central in the network as discussed in [8]. Therefore, sorting the concepts based on their weights should give a ranked list of concepts according to their importance. A representative subset of these concepts that are most central in the document is selected for each document. A similarity matrix is constructed for a pair of documents as shown in Table I, where the key concepts of each document are shown with the similarity score of every two concepts form each document.

The similarity score between concepts presented in Table I is fetched from the WordNet::Similarity WUP measure [17]. WUP measure calculates the relatedness of two concepts by measuring the depth of common ancestor of these concepts in the WordNet taxonomy. For our purpose, the maximum similarity score for each node is collected and the result is averaged. For example, in Table I the maximum similarity for concept "Internet" is 0.5 with the concept "subject", the same process is repeated for "network", "web", and "computer" and the results are averaged.

TABLE I
SEMANTIC-SIMILARITY MATRIX FOR DOCUMENT "INTERNET" AND DOCUMENT "KNOWLEDGE".

WUP	Internet	network	Web
Knowledge	0.25	0.54	0.54
understanding	0.22	0.46	0.46
subject	0.5	0.58	0.60

III. EXPERIMENT

A. Experiment design and evaluation

We applied our comprehension-based method in three cases: In case 1, we compare two unstructured documents with the same topic. This is one of the cases that can be missed in traditional vector based methods due to lack of term occurrence in short texts, therefore it holds special importance in our case. In case 2, we compare two documents with completely unrelated topics. This is important as a boundary case, to demonstrate the ability to classify related/unrelated documents. In case 3, we compare a document to itself, demonstrating the performance of our method in cases where documents do share the same terms.

Additionally, for the purpose of evaluating our method, we have constructed a survey with the same three previously mentioned cases. 25 human readers were asked to compare the semantic similarity and topic of documents, and we will use the survey results to evaluate the performance of our method and vector-based methods.

B. Material

The documents were not edited or modified in anyway, except for removing hyperlinks and references. The selection was from the top of the document, and each document had 30 to 37 concepts, and 10 to 16 sentences. Our method starts by extracting the concepts in each sentence, and the comprehension model converts each document into a weighted concept network representation. Figure 4 shows a sample concept network of the Wikipedia document Knowledge. The size of the concept represents the weight of that concept and the color of the concept represents acquisition time. This representation is unique to our method, as vector based methods neglect acquisition time and the semantic relations between concepts. Additionally, this representation allows for examining the metrics of semantic networks and their structure for semantic networks analysis [18]. Quantitative metrics such as graph density, average path length, and clustering coefficient can expose qualitative insights about the semantic network. For our purpose, we select the central cluster of concepts in the document semantic network as a representation of the document. This selection is based on the weight of the concept, where the concepts that are more central in a document hold a higher weight than concepts with less value. Additionally, we

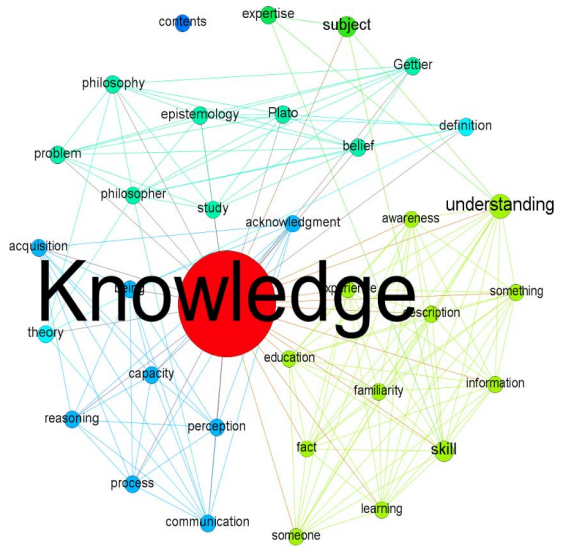

Fig. 4. Concept network representation of Wikipedia document "knowledge".

expect the concept with the highest weight in the document network to represent the topic of that document.

IV. RESULTS

A. Case 1: Comparing Two Documents with the Same Topic

For this case, we took two documents that share the same topic, but they do not share similar sentences or structure. The selected topic was knowledge, and the two documents were parts of the Wikipedia page and the Internet encyclopedia of philosophy page on knowledge.

TABLE II
SEMANTIC-SIMILARITY RESULTS FOR TWO TOPIC-RELATED DOCUMENTS USING COMPREHENSION-BASED METHOD AND VECTOR-BASED METHODS.

measure	Comprehension-based	Jaccard	cosine	dice
similarity score	0.84	0.12	0.18	0.18

The results of our method in comparison to vector-based similarity measures are presented in Table II. As mentioned previously, this is an important case where vector-based methods show the highest error, as the text is not long enough to establish the statistical regularity for term co-occurrence frequency. Our comprehension-based method evaluated the semantic similarity between the two documents at 84%, and it was able to identify the topic of the documents correctly based on the concept with the highest weight in the network representation.

B. Case 2: Comparing Two Documents with Unrelated Topics

For this case we took two documents with different topics. The selected topics were knowledge and Holi (festival from south Asia), and the two documents are from the Wikipedia pages on each topic. The same editing restrictions presented previously apply.

The semantic-similarity score results for two unrelated documents are presented in Table III. This case is important

TABLE III
SEMANTIC-SIMILARITY RESULTS FOR TWO UNRELATED DOCUMENTS USING COMPREHENSION-BASED METHOD AND VECTOR-BASED METHODS.

measure	Comprehension-based	Jaccard	cosine	dice
similarity score	0.25	0.07	0.10	0.10

for testing the performance of our method in classifying related/unrelated documents, and our method was able to identify the documents as unrelated (score less than 0.5).

C. Case 3: Comparing a Document to Itself

This is a special case demonstrating a comparison when the documents are identical. The similarity matrix is set to 1s diagonally, as demonstrated in Table IV. This is because any concept is similar to itself with the maximum value, one.

TABLE IV
A SAMPLE SIMILARITY MATRIX CONSTRUCTED USING THE ABOVE-AVERAGE KEY CONCEPT SELECTION METHOD FOR CASE 3.

WUP	Knowledge	understanding	subject	skill
Knowledge	1	0.79	0.83	0.83
understanding	0.76	1	0.79	0.8
subject	0.83	0.79	1	0.71
skill	0.83	0.8	0.71	1

TABLE V
SEMANTIC-SIMILARITY RESULTS FOR TWO IDENTICAL DOCUMENTS USING COMPREHENSION-BASED METHOD AND VECTOR-BASED METHODS.

measure	Comprehension-based	Jaccard	Cosine	Dice
similarity score	1	1	1	1

The Semantic-similarity results for two identical documents using comprehension-based method and vector-based methods are presented in Table V. This case is important to measure the performance of our comprehension-based method when word co-occurrence and frequency are significant to the comparison. Our method showed accurate results which are similar to the vector-based methods in this case.

D. Evaluation

We evaluate our comprehension-based method in two ways, first in comparison to vector-based methods and second in comparison to human performance from our survey. In comparison with vector based methods our framework was able to capture the semantic similarity where vector based methods fail, in case 1 as shown in Table VI. In case 2, our method provided a similarity score representing unrelated documents and so did vector-based methods. In case 3, the identical document was given a score of 1 with all methods, which is consistent with the notion of similarity.

When looking at the performance of our method in relation to human readers, we can see that our method provided a score of similarity that approximates that provided by human readers in case 1. And in case 2, our method provided a score representing dissimilarity although vector-based methods provided closer results to human readers. Finally, comparing

TABLE VI
COMPARING RESULTS FROM OUR FRAMEWORK WITH THREE VECTOR BASED METHODS.

	humans	Comprehension-based	Jaccard	Cosine	Dice
case 1	0.69	0.84	0.12	0.18	0.18
case 2	0.16	0.25	0.07	0.10	0.10
case 3	1	1	1	1	1

a document to itself provided an identical score to that from human readers in case 3.

The previous results are a proof of concept for a comprehension-based similarity method, and we plan to extend the experiments to include a much larger corpus of text. At the same time, our results highlight the importance of using cognitive based methods to solve complex information processing problems. Humans are naturally advantaged in identifying the similarity and topic of text documents, and our comprehension-based method mimics a human reader to perform semantic similarity and classification tasks. The advantage of our method is that it does not require a large corpus for training, and it does not treat the document as a bag of words. Instead, our method adopts a graph-model to represent documents and compare them. This gives us the advantage of looking at the intricate relations between semantic concepts and use graph-based analysis methods to analyze documents.

V. CONCLUSION

We presented a comprehension-based framework for measuring semantic similarity. The framework pays special attention to context, location, and acquisition time of each semantic concept in the document, unlike vector based methods. The weighted concept-network representation of a document has the advantage of minimizing redundancy, and that semantic relations between concepts can be analyzed, measured, and mined for insights about the document. Three experiments were presented with three cases: comparing a document to itself, comparing documents with the same topic, and comparing documents with unrelated topics. In addition, we compared our results with three popular vector-based similarity measures and human readers. The presented comprehension-based framework was able to successfully measure semantic similarity in situations where vector-based methods fail.

REFERENCES

[1] A. Huang, "Similarity measures for text document clustering," in *Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand*, 2008, pp. 49–56.

[2] D. Metzler, S. Dumais, and C. Meek, *Similarity measures for short segments of text*. Springer, 2007.

[3] P. D. Asanka, "Finding similar files using text mining," in *Computer Science & Education (ICCSE), 2013 8th International Conference on*. IEEE, 2013, pp. 431–435.

[4] F. M. Couto and H. S. Pinto, "The next generation of similarity measures that fully explore the semantics in biomedical ontologies," *Journal of bioinformatics and computational biology*, vol. 11, no. 05, p. 1371001, 2013.

[5] W. Maowen, Z. C. Dong, L. Weiyao, and W. Q. Qiang, "Text topic mining based on lda and co-occurrence theory," in *Computer Science & Education (ICCSE), 2012 7th International Conference on*. IEEE, 2012, pp. 525–528.

[6] G. Varelas, E. Voutsakis, P. Raftopoulou, E. G. Petrakis, and E. E. Milios, "Semantic similarity methods in wordnet and their application to information retrieval on the web," in *Proceedings of the 7th annual ACM international workshop on Web information and data management*. ACM, 2005, pp. 10–16.

[7] T. K. Landauer and S. T. Dumais, "A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge." *Psychological review*, vol. 104, no. 2, p. 211, 1997.

[8] N. S. Al Madi and J. I. Khan, "Is learning by reading a book better than watching a movie? a computational analysis of semantic concept network growth during text and multimedia comprehension," in *Neural Networks (IJCNN), 2015 International Joint Conference on*. IEEE, 2015, pp. 1–8.

[9] L. Huang, D. Milne, E. Frank, and I. H. Witten, "Learning a concept-based document similarity measure," *Journal of the American Society for Information Science and Technology*, vol. 63, no. 8, pp. 1593–1608, 2012.

[10] M. Hardas and J. Khan, "Concept learning in text comprehension," in *Brain Informatics*. Springer, 2010, pp. 240–251.

[11] M. Rafi and M. S. Shaikh, "An improved semantic similarity measure for document clustering based on topic maps," *arXiv preprint arXiv:1303.4087*, 2013.

[12] L. Verhoeven and C. Perfetti, "Advances in text comprehension: Model, process and development," *Applied Cognitive Psychology*, vol. 22, no. 3, pp. 293–301, 2008.

[13] W. Kintsch, "The role of knowledge in discourse comprehension: A construction-integration model," *Psychological Review*, vol. 95, pp. 163–182, 1988.

[14] ——, "The construction-integration model of text comprehension and its implications for instruction," *Theoretical models and processes of reading*, vol. 5, pp. 1270–1328, 2004.

[15] C. D. Manning, M. Surdeanu, J. Bauer, J. R. Finkel, S. Bethard, and D. McClosky, "The stanford corenlp natural language processing toolkit." in *ACL (System Demonstrations)*, 2014, pp. 55–60.

[16] Q. Wu, C. Zhang, X. Deng, and C. Jiang, "Lda-based model for topic evolution mining on text," in *Computer Science & Education (ICCSE), 2011 6th International Conference on*. IEEE, 2011, pp. 946–949.

[17] T. Pedersen, S. Patwardhan, and J. Michelizzi, "Wordnet:: Similarity: measuring the relatedness of concepts," in *Demonstration papers at HLT-NAACL 2004*. Association for Computational Linguistics, 2004, pp. 38–41.

[18] P. Drieger, "Semantic network analysis as a method for visual text analytics," *Procedia-Social and Behavioral Sciences*, vol. 79, pp. 4–17, 2013.