Paper ID: 324

Domain-Based Similarity Calculation Method for Calculating Document Similarity

HMLG Herath# and BTGS Kumara

Department of Computing & Information Systems, Faculty of Applied Sciences Sabaragamuwa University of Sri Lanka

hmlgherath.std.appsc.sab.ac.lk

Abstract: Document similarity is important in different areas dealing with textual data such as knowledge management, information extraction, natural language processing, and artificial intelligence. Several methods are existing to calculate document similarity. But the results of most approaches are unsatisfactory because specific domain and contextual similarity are not taken into consideration. In this paper, a domain-based similarity calculation method to calculate document similarity is proposed by integrating context, World Wide Web (WWW), and WordNet Similarity. Context is gathered by implementing a topic modeling algorithm and generating a domain context. There are many topic modeling algorithms available and here Latent Dirichlet Allocation (LDA) is used. The World Wide Web is used to capturing the latest knowledge. The method makes it possible to get a similarity value to the words in different domains. The quality of the obtained model is compared and evaluated using human judgment to ensure the accuracy of the calculation. Results indicate the accuracy of the calculation and the proposed model can achieve the limitations of existing measures.

Keywords: Domain-based Similarity, Topic modeling, Wordnet Similarity, World Wide Web

Introduction

A. Background

Information required by the users today become varies and users deal with textual data more than the numerical data (Niharika, Latha and Lavanya, 2012). Since most information (more than 80%) is stored as text, it is believed that text mining has a high commercial value (Korde, 2012). It is a difficult task to apply data mining techniques to textual data instead of numerical data (Hotho, Nürnberger, and Paaß, 2005). Text mining is the mining of textual data. Text mining or text data mining is the process of obtaining high-quality information from text by analyzing and exploring unstructured text data (Potdar and Pattewar, 2016)(Aggarwal and Zhai, 2013). In the Text-Mining similarity measure plays a major role (Feldman et al., 2009).

A similarity measure is a function that calculates the degree of similarity between a pair of objects. Information acquiring, text classification, document clustering, question generation, text summarization similarity measuring is some areas that play an important role in text-similarity measuring (Agnihotri, Verma and Tripathi, 2014). When considering document similarity there are already exists plenty of techniques and tools and novel researches have been produced almost every year. In those research papers, the aim has been to explore and evaluate different techniques for similarity measures. The fundamental part of text similarity is finding the similarity between words. Then it can apply to sentences, paragraphs, and document similarities (Huang et al., 2012).

If words have a similar character sequence, they are similar lexically. This estimates the degree of similarity between the word sets of two given languages. Semantic similarity

means the likeness among text and document based on their contextual meaning (H. Gomaa and A. Fahmy, 2013). Determining semantic similarity between word pairs is an important component of text understanding which enables textual resources to be processed, classified, and structured (Majumder, Pakray, Gelbukh and Pinto, 2020). Ontologies, thesauri, domain corpora are several approaches used in the past, for assessing word similarity (Batet, Sánchez, and Valls, 2011). WordNet is the most popular knowledge base, which has been widely applied in many studies.

In similarity measures, it maps the distance or similarity between the symbolic descriptions of two objects into a single numeric value. It depends on two factors; properties of the two objects and the measure itself. Traditional approaches represent documents as a bag of- words and compute document similarities using measures like cosine similarity, Jaccard, Pearson Correlation Coefficient, Averaged Kullback-Leibler Divergence, and dice. From the research findings, the results of some existing approaches of calculating document similarity are unsatisfactory, as much specific semantic knowledge and contextual similarity are not taken into consideration (Agnihotri, Verma and Tripathi, 2014). And some methods are not up to date knowledge, time-consuming, and need more expert knowledge. (Feng, Wei, Lu and Dang, 2014).

B. Motivation

The traditional methodologies used in calculating document similarity is focused and analyzed in this study. Cosine similarity, Jaccard similarity, Dice coefficient, Pearson correlation, TF-IDF, Clustering methods like K-means, k-medoids are some methods recently used. When considering the cosine similarity, it still can't handle the semantic meaning of the text perfectly (Rahutomo, Kitasuka, and Aritsugi, 2012). Implementing cosine similarity calculation between two-term vectors often produces syntactically inconsistent results. Syntax matching may not be able to meet the difference of the problem with semantic meaning. For further process, the information retrieval system, it may produce a false result and cause degrading in its performance (Rahutomo, Kitasuka and Aritsugi, 2012). Cosine similarity is not of up-to-date knowledge. That means if cosine similarity is implemented into a case, there exists a synonym relation or a hypernym-hyponym relation the similarity result is low (Madylova and Öğüdücü, 2009).

The most common method utilizes a lexical database as a semantic network is Wordnet (Sebti and Barfroush, 2008). The similarity between two concepts can be derived based on WordNet's exploration (Curran, 2003). WordNet glosses as a corpus of contexts one obtains about 1.4 million words, which should be processed to create the context vectors introducing a noticeable computational cost (WordNet, 2009). These measures perform poorly with some terms due to the limited coverage of specialized domains in the knowledge models (Batet, Sánchez, and Valls, 2011). They are manually created by knowledge experts, so, they represent the ideal context of a concept. WordNet does not include information about the pronunciation of words and it contains only limited information about usage. It aims to cover most of everyday English and does not include much domain-specific terminology (Kilgarriff and Fellbaum, 2000).

C. Problem Statement and Research Questions

Measuring the sentence similarity is useful in various research fields, such as artificial intelligence, knowledge management, and information retrieval. Many methods are existing to measure document similarity. But the results of the most traditional approaches are unsatisfactory, as much specific semantic knowledge, contextual

similarity and domain are not taken into consideration. And some methods are not up to date knowledge, time-consuming and need more expert knowledge.

The following are the research questions relative to this research.

RQ1: How we can integrate existing methods of calculating document similarity?

RQ2: What are the trending context-based and semantic similarity methods for calculating document similarity?

RQ3: How we can capture the context?

RQ4: What are the main issues and limitations of integrating existing methods of measuring similarity?

D. Research Objectives and Goals

According to the study, there are many types of research in calculating document similarity that has conducted using traditional methods and also using some hybrid methods. During the systematic literature review, several gaps and shortcomings have been found in the existing researches. The objective of this work is to address those limitations by proposing a novel methodology.

1) Main Objective:

Propose a domain-based similarity calculation method for calculating document similarity

2) Specific objectives:

Capture the latest knowledge of the area of calculating document similarity.

Integrate the existing methods of measuring similarity.

Generating a domain context.

Methodology

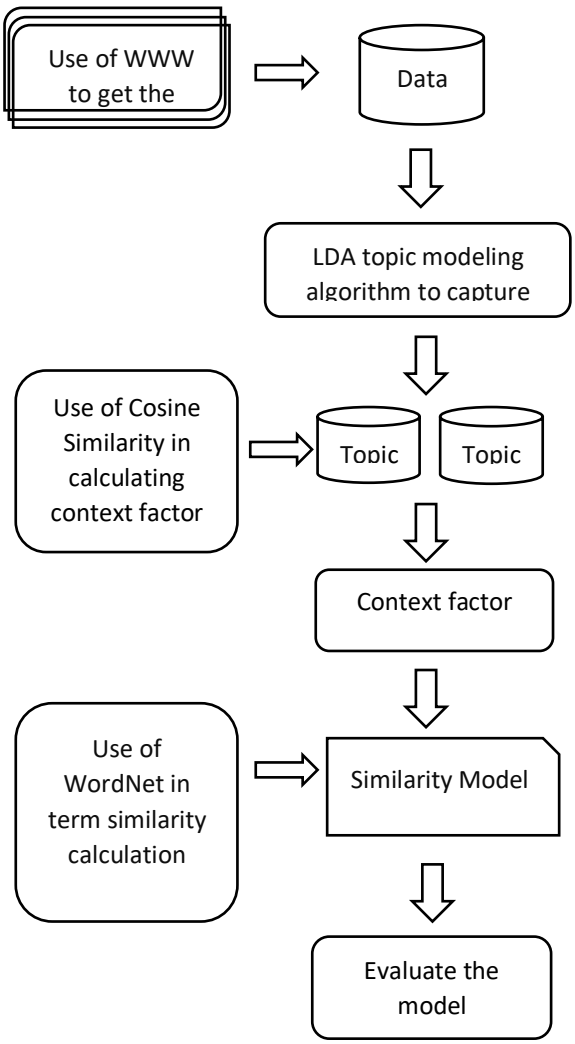

Figure 1. Methodology

A. Data set and Requirement analysis

Input for this research is collected using the World Wide Web. The World Wide Web is the universe of network-accessible information, a representation of human knowledge. It connected the world in a way that was not possible before and made it much easier for people to get information, share, and communicate. It provides increased capacity, greater sample diversity, easier access and convenience, lower costs, and time investment in data collection. The main intend to use WWW in here is to capture the latest knowledge captured using the WWW.

In a natural language environment, the semantic similarity of the same pair of words

may have a large difference in different areas of knowledge (Feng, Wei, Lu and Dang, 2014). As the initial step, a list of words is selected. For example, when considered the words 'Apple' and the 'Computer', apple is a fruit and a computer is a machine. It will give a less similarity value. But if we consider the two words in the technology domain, apple is a type of computer. There should be a high similarity value then. As above mentioned, a list of word pairs representing two domains is selected. Individual search results of each word and combined search results of the words have been gathered. There are unlimited search results for input in the World Wide Web. So, the amount of data captured is limited here. The data is collected manually and the collected data are saved into CSV files.

Table 1. List of words used

Word1	Word2	Combined Results
Apple	Computer	Apple Computer
Apple	Fruit	Apple Fruit
Blackberry	Fruit	Blackberry Fruit
Blackberry	Phone	Blackberry Phone
Bat	Animal	Bat Animal
Bat	Game	Bat Game
May	Tree	May Tree
May	Month	May Month
Orange	Color	Orange Color
Orange	Fruit	Orange Fruit
Ambulance	Vehicle	Ambulance Vehicle
Ambulance	Medical	Ambulance Medical
Minute	Time	Minute Time
Minute	Report	Minute Report
Ball	Dance	Ball Dance
Ball	Game	Ball Game
Bank	Money	Bank Money
Bank	River	Bank River
Band	Music	Band Music
Band	Hair	Band Hair
Gold	Jewelry	Gold Jewelry
Gold	Color	Gold Color
Windows	OS	Windows OS
Windows	Glass	Windows Glass
Capital	Letter	Capital Letter
Capital	Country	Capital Country
Wheel	Vehicle	Wheel Vehicle
Wheel	Tire	Wheel Tire
Ring	Bell	Ring Bell
Ring	Jewelry	Ring Jewelry
Nail	Finger	Nail Finger
Nail	Pin	Nail Pin
Palm	Tree	Palm Tree
Palm	Hand	Palm Hand
Rock	Music	Rock Music
Rock	Stone	Rock Stone
Rose	Flower	Rose Flower
Rose	Color	Rose Color
Toast	Drink	Toast Drink
Toast	Food	Toast Food

B. Implementing the Latent Dirichlet Allocation (LDA) topic modeling algorithm

Topic modeling is an unsupervised machine learning technique that is capable of scanning a set of documents, detecting word and phrase patterns within them, and automatically clustering word groups and similar expressions that best characterize a set of documents. Latent Dirichlet Allocation (LDA) is an example of a topic model and is used to classify text in a document to a particular topic. It builds a topic per document model and words per topic model, modeled as Dirichlet distributions (Topic Modeling using Latent Dirichlet Allocation (LDA) | Honing Data Science, 2020). In this research, context is gathered with the use of LDA. Input to the LDA is the collected data from the Web. Then the topics for each input file are gathered. The LDA was applied in python using Jupyter Notebook. The process followed is mentioned below.

1) Loading data

Captured data from the World Wide Web is used as the input data set to LDA. Each CSV data file contains the search results for the selected words in Table 1. Then that CSV file was read.

2) Data cleaning

A simple preprocessing was done on the content of files to make them more amenable for analysis, and reliable results. To do that, a regular expression is used to remove punctuation, and then lowercase the text.

3) Exploratory analysis

To verify whether the preprocessing happened correctly, a wordcloud is made using the wordcloud package to get a visual representation of most common words. It is key to understanding the data and ensuring that whether the right track is followed, and if any more preprocessing is necessary before training the model.

4) Preparing data for LDA analysis

The next, step is to transform the textual data in a format that will serve as an input for the training LDA model. It is started by converting the documents into a simple vector representation (Bag of Words BOW). Next, convert a list of titles into lists of vectors, all with length equal to the vocabulary. Then plot the ten most frequent words based on the outcome of this operation (the list of document vectors). As a check, these words should also occur in the word cloud.

5) LDA model training and results visualization

Then the number of topic parameters need is input and visualized the results.

C. Implementing Cosine Similarity

Cosine Similarity is a measure of similarity that can be used to compare documents and give a ranking of documents concerning a given vector of query words. In this research, the Cosine similarity algorithm was used which is implemented in Java as the programming language and Eclipse is the IDE. In this case, input to the cosine similarity is the topics resulted from the LDA, any value between 0 and 1 is given as the output of the algorithm.

D. Calculating WordNet Similarity

WordNet is also going to consider creating the similarity model which is an ontology-based similarity calculation method. It is a large lexical database of English. Nouns, verbs, adjectives, and adverbs are grouped into sets of cognitive synonyms or synsets, each expressing a distinct concept where synsets are interlinked through conceptual-semantic and lexical relations In the initial step of collecting data, a list of domain-based words is selected. Here, WordNet is used to compare and get the similarity of that selected pair of words.

If it is describing further, as the initial step researcher installs the WordNet software to the personal computer. After installing WordNet, an algorithm was implemented for calculating the similarity of words.

E. Integrating Methods

The final part of the methodology is to integrate the implemented methods. There a novel equation is created as follows.

Term Similarity= Wordnet Similarity + Context factor

The gathered results of the WordNet similarity value and the Cosine similarity values are applied in the above equation and get a value for each pair.

Results

The initial part of the research is to collect data using the World Wide Web. The selected word pairs as stated above, considering their domain are represented in Table 1. There are unlimited search results to

a word. So here search results gathered are limited, and the search results of each word and the word pairs were gathered.

Then that data set of search results were applied to the topic modeling algorithm LDA and gathered the topics of each word. Then those topics were preprocessed and calculated the cosine similarity of them.

At the same time, the WordNet similarity of the selected word pairs was calculated. In the initial part, a pair of domain-based words were selected. The WordNet similarity of that pairs is calculated here using the implemented algorithm. As a result, a value between 0 and 1 has resulted.

The next step of integrating the methods is processed using the cosine similarity results and the WordNet similarity results. There the results were applied to the following equation and calculated a term similarity value for each pair.

Term Similarity= WordNet Similarity + Context factor

After the calculation of term similarity, it was evaluated using human judgment. They were asked to give a similarity value for each pair of words. Then calculated the average similarity value for each pair of words.

The Pearson correlation was used to compute the correlation between human ratings and the term-similarity methods and calculated as:

$$r(HR, IRT)= \frac{\sum(ts_1 ts_2)-\frac{\sum ts_1 \sum ts_2}{N}}{\sqrt{\prod_{i=1}^{2}(\sum ts_i^2-\frac{(\sum ts_i)^2}{N})}}$$

Here, HR is the human rating and IRT is the Information retrieval-based term-similarity method. Parameters ts1 and ts2 are the human-rating similarity and the term similarity for terms, respectively. Parameter N is the number of term pairs.

The coefficient value can range between -1.00 and 1.00. If the coefficient value is in the negative range, then that means the relationship between the variables is negatively correlated, or as one value increases, the other decreases. If the value is in the positive range, then that means the relationship between the variables is positively correlated, or both values increase or decrease together.

Table 2. Correlation between human rating and term-similarity methods

Term1	Term2	Context-Aware Similarity	Human Rating	Wordnet Similarity
Apple	Computer	1	0.57	0.06
Ambulance	Vehicle	0.8	0.78	0.6
Blackberry	Fruit	1	0.73	0.3
Bat	Animal	1	0.69	0.5
May	Month	0.81	0.86	0.76
May	Tree	0.76	0.63	0.6
Minute	Time	0.99	0.85	0.89
Minute	Report	0.54	0.6	0.75
Ball	Dance	1	0.72	0.93
Bank	Money	0.68	0.85	0.85
Band	Music	1	0.81	0.68
Band	Hair	0.81	0.66	0.57
Gold	Jewelry	1	0.84	0.11
Gold	Color	0.95	0.6	0.35
Windows	OS	1	0.83	0.92
Windows	Glass	1	0.64	0.6
Capital	Letter	0.9	0.82	0.83
Wheel	Tire	1	0.68	0.67
Ring	Jewelry	1	0.84	0.83
Nail	Finger	1	0.85	0.8
Palm	Hand	0.8	0.76	0.77
Rock	Music	1	0.77	0.69
Rock	Stone	1	0.8	0.83
Rose	Flower	0.81	0.86	0.82

Toast	Food	0.76	0.81	0.66
Correlation		0.82	1	0.5

The value of 0.82 has resulted in the Correlation between human rating and context-aware similarity and the value of 0.5 has resulted in the wordnet similarity and human rating. That means there is a positive relationship between those pairs of variables. For a positive increase in one variable, there is also a positive increase in the second variable. Here, both correlations are positive values but the context-aware similarity method has the highest correlation value compare to the WordNet method. That means the strength of the relationship between context-aware similarity and human rating is higher compared to the wordnet.

Discussion and Conclusion

In this paper, a domain-based similarity calculation method is proposed for calculating document similarity. Though there are many methods available to measure document similarity, many of them did not take into account the domain knowledge. The novel method is proposed by integrating existing methods to overcome the limitations of them. Context, World Wide Web (WWW), and Wordnet are considered here. The context is captured by implementing an LDA, a topic modeling algorithm. The World Wide Web is used to capture the latest knowledge. Methods were calculated separately and then integrated them by introducing a term similarity equation. Finally, the values were evaluated using human judgment. The experimental results (Table 2) show that the context-aware similarity method has the highest correlation value compare to the WordNet method which confirms that the relationship between the variables is positively correlated, and there is a higher relationship between context-aware similarity and human rating compared to the wordnet.

References

ggarwal, C. C. and Zhai, C. X. (2013) Mining text data, Mining Text Data. doi: 10.1007/978-1-4614-3223-4.

Agnihotri, D., Verma, K. and Tripathi, P. (2014) 'Pattern and cluster mining on text data', in Proceedings - 2014 4th International Conference on Communication Systems and Network Technologies, CSNT 2014. doi: 10.1109/CSNT.2014.92. Batet, M., Sánchez, D. and Valls, A. (2011) 'An ontology-based measure to compute semantic similarity in biomedicine', Journal of Biomedical Informatics. doi: 10.1016/j.jbi.2010.09.002.

Curran, J. (2003) 'From distributional to semantic similarity', University of Edimburgh. doi: 10.1.1.10.6068.

Feldman, R. et al. (2009) 'Introduction to Text Mining', in The Text Mining Handbook. doi: 10.1017/cbo9780511546914.002.

Feng, X., Wei, J., Lu, W. and Dang, J., 2014. Word Semantic Similarity Calculation Based on Domain Knowledge and HowNet. TELKOMNIKA Indonesian Journal of Electrical Engineering, 12(2).

H.Gomaa, W. and A. Fahmy, A. (2013) 'A Survey of Text Similarity Approaches', International Journal of Computer Applications. doi: 10.5120/11638-7118.

Hotho, A., Nürnberger, A. and Paaß, G. (2005) 'A Brief Survey of Text Mining', LDV Forum - GLDV Journal for Computational Linguistics and Language Technology. doi: 10.1111/j.1365-2621.1978.tb09773.x.

Honing Data Science. 2020. Topic Modeling Using Latent Dirichlet Allocation (LDA) | Honing Data Science. [online] Available at: <https://honingds.com/blog/topic-modeling-latent-dirichlet-allocation-lda/>

Huang, L. et al. (2012) 'Learning a concept-based document similarity measure', Journal of the American Society for Information Science and Technology. doi: 10.1002/asi.22689.

Kilgarriff, A. and Fellbaum, C. (2000) 'WordNet: An Electronic Lexical Database', Language. doi: 10.2307/417141.

Korde, V. (2012) 'Text Classification and Classifiers:A Survey', International Journal of Artificial Intelligence & Applications. doi: 10.5121/ijaia.2012.3208.

Madylova, A. and Öğüdücü, Ş. G. (2009) 'A taxonomy based semantic similarity of documents using the cosine measure', in 2009 24th International Symposium on Computer and Information Sciences, ISCIS 2009. doi: 10.1109/ISCIS.2009.5291865.

Majumder, G., Pakray, P., Gelbukh, A. and Pinto, D., 2020. Semantic Textual Similarity Methods, Tools, And Applications: A Survey.

Machine Learning Plus. 2020. Cosine Similarity - Understanding The Math And How It Works? (With Python). [online] Available at: <https://www.machinelearningplus.com/nlp/cosine-similarity/>

Medium. 2020. Topic Modeling And Latent Dirichlet Allocation (LDA) In Python. [online] Available at: <https://towardsdatascience.com/topic-modeling-and-latent-dirichlet-allocation-in-python-9bf156893c24>.

Niharika, S., Latha, V. S. and Lavanya, D. R. (2012) 'A survey on text categorization', International Journal of Computer Trends and Technology.

Potdar, D. S. and Pattewar, T. M. (2016) 'A novel similarity measure technique for clustering using multiple viewpoint based method', in Proceedings of the 10th International Conference on Intelligent Systems and Control, ISCO 2016. doi: 10.1109/ISCO.2016.7727007.

Rahutomo, F., Kitasuka, T. and Aritsugi, M. (2012) 'Semantic Cosine Similarity', Semantic Scholar.

Sebti, A. and Barfroush, A. A. (2008) 'A new word sense similarity measure in wordnet', in Proceedings of the International Multiconference on Computer Science and Information Technology, IMCSIT 2008. doi: 10.1109/IMCSIT. 2008.4747267.

WordNet (2009) 'WordNet Domains', WordNet Affect.

Acknowledgement

This research was supported by the Department of Computing & Information Systems, Faculty of Applied Sciences, Sabaragamuwa University of Sri Lanka.

Author Biographies

H. M. L. G. Herath an undergraduate student in the Sabaragamuwa University of Sri Lanka and will be graduating in 2020 with a BSc special in Information Systems.

Banage T. G. S. Kumara received the Bachelor's degree in 2006 from Sabaragamuwa University of Sri Lanka. He received the master's degree in 2010 from the University of Peradeniya and Ph.D degree in 2015 from the University of Aizu, Japan. His research interests include semantic web, web data mining, web service discovery, and composition.