The Implementation of using Medical ontologies in Plagiarism detection

K. Omar, B. Alkhatib, M. Dashash, F. Alhassan

Abstract – *This paper aims to present a new algorithm in plagiarism detection using semantic web tools and notions. For increasing detection accuracy we suggest using domain ontology in addition to global semantic resources. Using global semantic resources will increase the effect of ambiguity therefore we suggest using disambiguation techniques. Not all semantically similar texts are plagiarized. So, we suggested that another detection technique should be used in order to reduce false positive results. Although our work has done in medical domain and English language, it presents a generic algorithm that can be adapted for different domains and languages. For medical domain, a set of medical ontologies was used for enriching extracted medical terms. In the other side, WordNet was used for enriching global terms. The test results of the algorithm shows that it was able to detect advanced types of plagiarism that are out of the reach of classical methods such as: using word synonyms ,word re-ordering ,text re-styling and other natural languages techniques which are usually used to hide the plagiarism action.*

Keywords: *semantic web, medical ontologies, Plagiarism detection in medical Sciences*

I. Introduction

Plagiarism is the using of others ideas or publications without getting the permission from the work owner [1] and we can define plagiarism as following:

"… Passing off someone else's work, either intentionally or unintentionally, as your own, for your own benefit." [2]. A lot of research has been done in plagiarism detection in many languages, and many systems have been developed to detect plagiarism, but every one of these systems uses a different algorithms and different heuristics to detect and discover plagiarism, which is not easy problem because when someone plagiarizes someone else work he does his best to hide his manner by using Natural language features to re-style or re-explain others work to make this work belongs to him. So there are many problems face the plagiarism detection algorithms and systems, but the most important one is that when the plagiarist tries to understand the idea of the original text and reform it using his own words and writing style.

In response for that we try to track back this advanced plagiarism process and specify a number of common semantic features between the original text and the plagiarized text. Our investigations lead us to a new algorithm that exploits semantic similarity for detecting plagiarism.

This paper is structured as following: related works, proposed semantic algorithm for plagiarism detection, system design and implementation, tests and results, discussion, conclusion and future work.

II. Related Work

Like in most information processing domains, the extension of classical trends in plagiarism detection by semantic notions and tools has become indispensable. However, in the past few years, only a few numbers of related researches have been published in this concern. Almost all publications try to represent textual content by some semantic structure then use semantic measures for detecting non-classical plagiarism.

Some approaches represent text as graph of sentences where edges reflect semantic connection between sentences. Then a graph matching algorithm is used for computing similarity [20]. Another approach use ontology extraction techniques for representing texts as ontology then use ontology mapping for detecting similarity [21]. In more popular approach, semantic resources like WordNet and global ontologies is used for computing similarity between sentences depending on semantic distance between their words [22].

As we believe, what hinders exploiting semantic web in plagiarism detection is accuracy for that, tile today, semantic web still lacks accuracy and, in applications such as plagiarism detection, accuracy is nonnegotiable.

Manuscript received January 2007, revised January 2007

In response, we return the lack of accuracy to several reasons: firstly and most important, all suggested approaches use global ontology and semantic resources and it doesn't make use of domain ontology which is a powerful choice for enhancing accuracy in semantic web.
Secondly, using global ontology increases ambiguity effect which needs effective techniques for disambiguation.
Lastly, all semantic approaches will be very useful for detecting semantic similarity but not all semantic similar texts are plagiarized. Therefore, to get its best in plagiarism detections, we think that it is necessary for semantic techniques to be coupled with other plagiarism detections techniques such as citation based detections [4].

III. Proposed Semantic Method for Plagiarism Detection

Our proposed algorithm for plagiarism detection composes of two phases: semantic analyzing, semantic comparison. In this section we will give a brief description of the algorithm.

Semantic Analyzing

In this phase we aim at getting semantic representation of the text using global semantic resources like WordNet [6] and domain ontologies. As domain ontology we used EMBL-EBI an ontology lookup service which provides a unified service for about 93 medical ontology such as Geno Ontology (GO), Infectious Disease Ontology (IDO) and Foundational Model Of Anatomy Ontology (FMA) [7] [8].

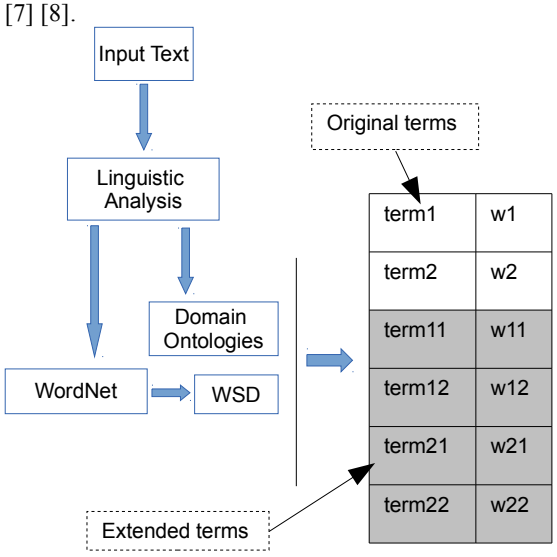


Figure 1 semantic analysis

Before extracting semantics the text goes in a linguistic preparation stage including: sentences segmentation using Stanford NLP tool [9], stemming and stop words removal [10], part of speech tagging using Stanford POS [11] and composed nouns extraction using Stanford dependencies parser [12]. Besides, we implemented a simple algorithm for extracting complete partitions such as "(*), "*" or [*]" which, most likely, represents some abbreviation to a domain term and is plagiarized as it is.
On this version of our work we chose to represent the text semantics in a simple vector of terms; then, these terms will be extended using domain ontologies and WordNet. At the end of the extension stage, we will get a rich vector of weighted terms as a representation of text semantics. As previously noted, using global semantic resources like WordNet for extension can add lots of unrelated concepts since any word can belong to different synsets. Therefore, we have implemented an adapted version of LESK algorithm for disambiguating words senses [24].
Depending on the comparing study provided by [27] we chose the spreading technique, for terms extension, which use several iterations to extend terms vector by related terms using ontology. In each iteration, all terms are inspected and related terms are appended. We use WordNet and domain ontology for retrieving related terms in each iteration.
Spreading terminates at one of these conditions:

- There is no other related terms left.
- Iteration number exceeds iteration threshold.
- Similarity result in the previous iteration is larger than current result.

As an iteration threshold we chose at most five iterations depending on the results and conclusion of [27].

Semantic comparison

After extracting semantic representations of the suspicious text and the target text, semantic comparison goes through two stages: Domain comparison and similarity measure.
Domain comparison
It is very clear that if the two texts are from different domains, then any further investigations will be meaningless.
In response, we defined a new function that computes domain closeness of two texts depending on their semantic representation as flows:

$$F(SR_{t_1}, SR_{t_2}) = \max_{O \in domainontdogies} \left\{ \sum_{c \in (O \cap SR_{t_1} \cap SR_{t_2})} W_{SR_{t_1}}(c) * W_{SR_{t_2}}(c) \right\}$$

SR_{t_i} Is a semantic representation of text *i.*

$W_{SR_{t_i}}(c)$ Is a function returning the weight of concept c in SR_{t_i} .

$c \in (O \cap SR_{t_1} \cap SR_{t_2})$ Are all shared concepts/terms between the two texts and domain ontology *O.*

Consequently, if domain comparison doesn't exceed closeness threshold the comparison stop at this stage and texts stated as not plagiarized.

Similarity Measure

Although there are big number of proposed researches concerning semantic similarity, we have found that almost all approaches can be classified into two categories: cosine similarity model [25] [26] [27] and graph model [20] [22] [27]. In [25] they extends terms vector by Is_a relation and compute the cosine of the two extended vectors; likewise, in [26] they used terms vector as semantic representation; then, they unify vectors dimensions using what they called dimensions equalization by semantic relations.

In the other side, in [20], [22] and [27] the two texts are represented as one bipartite graph where word similarity measure is used as edges weight. Afterward, they use graph matching algorithm to get best match between the two texts.

As shown in [27], cosine similarity measure with spreading by at most five iterations gives the best overall results. Thus, we chose to implement the same measure but with two main differences: using domain ontology (in addition to WordNet) for extension, and using word sense disambiguation.

IV. System Design and Implementation

Our developed system contains five main modules

-Searcher: we used "Bing" search API[15] the searcher input is a user manually put query which related to the file which the user want to detect plagiarism into it.

-Downloader: which download results sets of the user query, the downloaded files are stored into system database.

-Semantic analyzer: analyze the results sets semantically using medical ontologies of EMBL-EBI project API which support a lot of procedures for getting terms metadata and terms parents, and terms chils, and terms belonging ontology, and term relations, the semantic analyzer after getting the concepts from texts it stores these concepts into system database to build a knowledge base about concepts to use them instead of EMBL-EBI API project, that the system first search about the concept into system database and if it dosenot find it in the system database it search for it using EMBL-EBI API.

-Semantic Comparer: this module is compare between analyzing files concepts in a semantic manner not in an ordinary or string based manner.

-Report Viewer: this module views the semantic plagiarism detection which has already done by the semantic comparer.

All modules has programmed using java programming language using NetBeans IDE 8.0[16] as an environment for programming, and we used SQL Server 2008 DataBase to store the results sets into it.

As shown in the following figure below:

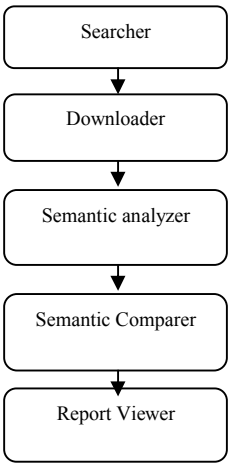


Fig. 4 System Structure

V. Tests and Results

the algorithm was tested on 100 abstracts from Europe PMC[17] web site which you can search On Europe PMC you can search all the content (abstracts and article full text) in a single search, whereas PubMed[18] and PMC are separate resources, and for every abstract we plagiarized it manually, the input files are separated as the abstracts count that for every abstract there is a file contains plagiarized texts parts and non plagiarized texts parts , the goal from testing the algorithm against non plagiarized text is to measure the false detection results of the algorithm, the evaluation is done according to the precision measure which are the described as the following formula [19] :

$$\text{precision} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{retrieved documents}\}|} \quad (2)$$

In plagiarism detection systems the precision is the number of correct plagiarism detection results divided by the number of all returned results.

In our developed system we test three versions of the developed algorithms:

-Plagiarism detection algorithm without expanding non – ontology concepts: in this version of developed algorithm the algorithm compared only between medical ontology concepts and public knowledge ontology concepts, the precision was about 85% that the analyzing of the text (also medical papers) contains a bout of 20% of terms that they are not belong to any medical ontology of public knowledge ontology.

-Plagiarism detection algorithm without disambiguation of public knowledge concepts: this version of developed algorithm its precision was about 83% that using of concepts without applying disambiguation generates false

detection results because of the algorithm compares between all public knowledge concepts synsests and this leads to occurrence of intersected words between concepts synsets.

-Plagiarism detection algorithm with disambiguation of public knowledge concepts and with non-ontology expanding: : in this version of developed algorithm the algorithm compared all concepts and used disambiguation heuristic, by using these tow power points the precision raised to 90% .

VI. DISCUSSION

Our developed algorithm (with its versions) has a lot of positive points in the view of plagiarism detection algorithms:

- Analyses the text in semantic manner by expanding and enrich terms with medical ontolgies and public knowledge ontology and generating concepts-profiles for non-ontology concepts.
- Compare between files through tow levels this first one is comparing to detect if the two files are belong to the same ontolgies, if so then the algorithm complete in comparing between all files concepts
- The using of disambiguation is giving the algorithm a very power full point for non-giving a false plagiarism detection results.
- The developed algorithm build and update a concept knowledge base by saving all concepts information into system local database and this decrease the global execution time of the algorithm (because retrieving data from local database is faster than retrieving it by using an online web API) .
- The algorithm is applicable to detect plagiarism in any language by providing it with the required ontolgies.

VII. Conclusion and future work

Our developed system with its algorithm versions, from our point of view is an important step and technology in plagiarism detection systems, and our future mission is to develop this system to become full language in-depended by providing it with the ontolgies for other language(especially for Arabic language).

Acknowledgements

This study is supported by Damascus University, Syria

References

[1] Vinod K.R.,Sandhya.S,Sathish Kumar D,Harani A,David Banji and,Otilia JF Banji , "Plagiarism history ,detection and prevention ", Hygeia*: journal for drugs and medicines, Vol.3*-Issue.1-pp. 1- 4, 2011

[2] Carroll, J. (2002) *A Handbook for Deterring Plagiarism in Higher Education. Oxford: Oxford Brookes University*

[3]Tachaphetpiboon, S.; Facundes, N.; Amornraksa, T., Plagiarism indication by syntactic-semantic analysis, *Asia-Pacific Conference on Communications*, pp.237-240, 2007

[4] Gipp, B., & Beel, J. (2010)." Citation based plagiarism detection - A new approach to identify plagiarized work language independently". HT'10 - *Proceedings of the 21st ACM Conference on Hypertext and Hypermedia*, (June), 273–274.

[5]The OBO Foundry .[online] Available at: http://www.obofoundry.org/ [Accessed 15-March 2015].

[6] Princeton University, WordNet a large lexical database of English.[online] https://wordnet.princeton.edu [Accessed 15-March 2015].

[7] EMBL-EBI, Ontology Lookup Service.[online] http://www.ebi.ac.uk/ontology-lookup/init.do#soft [Accessed 15-August 2015].

[8] EMBL-EBI, Ontology Lookup Service.[online] http://www.ebi.ac.uk/ontology-lookup/ontologyList.do [Accessed 15-August 2015].

[9] The Stanford Natural Language Processing Group.[online] http://nlp.stanford.edu/software/ [Accessed 15-August 2015].

[10] Wikipedia,the free encyclopedia.[online] https://en.wikipedia.org/wiki/Stop_words / [Accessed 20-August 2015].

[11] The Stanford Natural Language Processing Group. Stanford Log-linear Part-Of-Speech Tagger [online] http://nlp.stanford.edu/software/tagger.shtml/ [Accessed 1-August 2015].

[12] The Stanford Natural Language Processing Group. Stanford Dependencies [online] http://nlp.stanford.edu/software/stanford-dependencies.shtml/ [Accessed 1-August 2015].

[13] Abrate, M., Bacciu, C., Marchetti, A., & Tesconi, M. (2012). WordNet Atlas: a web application for visualizing WordNet as a zoomable map.

[14] Wikipedia, Cosine similarity.[online] https://en.wikipedia.org/wiki/Cosine_similarity/ [Accessed 1-August 2015].

[15] Bing, API Basics. [online] Available at: http://www.bing.com/developers/s/APIBasics.html [Accessed 15-March 2015].

[16] NETBEANS , [online] Available at: https://netbeans.org/ [Accessed 1-March 2014].

[17]Europe PMC , [online] Available at: https://europepmc.org/ [Accessed 1-March 2014].

[18] PubMed.gov, [online] Available at: http://www.ncbi.nlm.nih.gov/pubmed [Accessed 20-March 2014].

[19] Wikipedia, Precision and recall.[online] https://en.wikipedia.org/wiki/Precision_and_recall/ [Accessed 1-August 2015].

[20] A. H. Osman, N. Salim, M. S. Binwahlan, H. Hentably, and A. M. Ali, "Conceptual Similarity and Graph-Based Method," Journa Theoretical and Applied Information TechnologyOctober, vol. 32

[21] M. Shenoy, K.C.Shet, and U. D. Acharya, "S Emantic P Lagiarism Detection System Using Ontology Mapping," Advanced Computing, vol. 3, no. 3, pp. 59–62, 2012

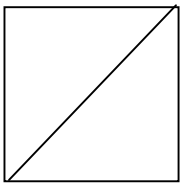
[22] Y. Palkovskii, A. Belov, and I. Muzyka, "Using WordNet-based Semantic Similarity Measurement in External Plagiarism Detection - Notebook for PAN at CLEF 2011", in Proc. CLEF (Notebook Papers/Labs/Workshop), 2011.

[23] Salha Alzahrani Naomie Salim "Fuzzy Semantic Based String Similarity for Extrinsic Plagiarism Detection" Lab Report for PAN at CLEF 2010

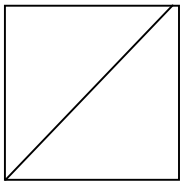
[24] Satanjeev Banerjee, Ted Pedersen An Adapted Lesk Algorithm for Word Sense Disambiguation Using WordNet

[25] MADYLOVA, A, OGUDUCU, S.G. A TAXONOMY BASED SEMANTIC SIMILARITY OF DOCUMENTS USING THE COSINE MEASURE
[26] FAISAL RAHUTOMO TERUAKI KITASUKA, SEMANTIC COSINE SIMILARITY.
[27] RAJESH THIAGARAJAN, GEETHA MANJUNATH, AND MARKUS STUMPTNER HP LABORATORIES HPL-2008-87 "COMPUTING SEMANTIC SIMILARITY USING ONTOLOGIES"

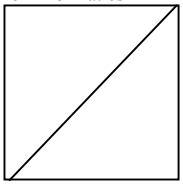
Authors' information



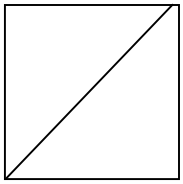
Damascus University, Syrian Arab Republic. K. Omar, PhD student in Faculty of informatics Engineering, Damascus University, Syria, Master degree in artificial intelligence, Damascus university, Syria. He obtained his Master degree in artificial intelligence about plagiarism detection, and he is currently a PhD student undertaken research on plagiarism detection using semantic web techniques. Eng. Omar is a member of Syrian Scientific Society for Informatics



B. AlKhatib, An associate professor at Damascus University, Faculty of Informatics Engineering, Damascus University, Syria, PhD in Artificial intelligence, France, Director of (MWT , MWS) programs in the Syrian virtual university , Director of the Syrian-COMSATSCOMSTECH IT Centre (SCCITC) Damascus, Syria. He has many international publications related to web data mining, plagiarism detection, natural language processing, and virtual learning. Dr. AlKhatib is a member of Syrian Scientific Society for Informatics



M. Dashash, An associate professor in the Faculty of Dentistry, Damascus University, Syria, Director of Evaluation and accreditation, Ministry of Higher Education, Syria. She has her Doctor of Dental Surgery DDS, from Damascus University, Syria, Master degree in Peadiatric Dentistry Msc, from Damascus University, Syria, PhD in Peadiatric Dentistry, from University of Manchester, United Kingdom, and Membership of the Faculty of Dental surgery, Royal College of Surgeon of Edinburgh MFDS RCS(Ed), from United Kingdom, She has 23 international publications related to dentistry, public health, health informatics, genetics, quality assurance, Medical education, curriculum reform, and electronic databases. Dr. Dashash is a member of the Syrian Dental Association, Alumni Association of University of Manchester, and a member of the Royal College of Surgeon of Edinburgh, United Kingdom



B. AlKhatib, An associate professor at Damascus University, Faculty of Informatics Engineering, Damascus University, Syria, PhD in Artificial intelligence, France, Director of (MWT , MWS) programs in the Syrian virtual university , Director of the Syrian-COMSATSCOMSTECH IT Centre (SCCITC) Damascus, Syria. He has many international publications related to web data mining, plagiarism detection, natural language processing, and virtual learning. Dr. AlKhatib is a member of Syrian Scientific Society for Informatics