PATENTS DUPLICATE DETECTION USING SEMANTIC APPROACH IN MEDLINE

S. Sridharani[1], J. Ganesh[2]

[1]Student Member, Department of Computer Science and Engineering,

[2] Staff Member, Department of Computer Science and Engineering,

Arasu Engineering College, Kumbakonam, TamilNadu, India.

Abstract: Plagiarism detection is the process of locating instances of plagiarism within a work or document. Information retrieval is widely applied in indentifying similarity between two text/document/concepts. In this paper, to compare the patents similarity are identify in the MEDLINE. A patent holds a wide detail of information such as descriptions and claims. To achieve this, first preprocessing stage is handled to remove the stop words. After that, to find out the stemming words by using suffix and prefix stripping algorithm. Second stage of query expansion is to improve the retrieval performance in IR operation, finding all the various morphological forms of words by using Wu-Palmer algorithm. Wu-Palmer is used to measure the similarity between two sources(Senticnet and Wikitionary). Its calculating the similarity by considering the depths of the two concept in the UMLs. The proposed work is to gives the rights to stop others from copying, manufacturing, selling or importing others work.

Keyterms: Information Retrieval, Plagiarism Detection, MEDLINE, UMLS Metathesarus, Word Sense Disambiguation, Wu-Palmer.

I. INTRODUCTION

Information retrieval, as the name implies, concerns the retrieving of relevant information from databases. It is basically concerned with facilitating the user's access to large amounts of (predominantly textual) information.Academic area Plagiarism is major problem to copying of someone else information. It has two stages the first stage is Intentional plagiarism- the candidate should known the contents which are taken from some other authors. The second stage is Unintentional plagiarism- the candidate doesn't known which is taken from others documents. The plagiarism detection should improve the student knowledge and then they can learn about the several fields.

MEDLINE (Medical Literature Analysis and Retrieval System Online) is a large collection of documents. The academic journals are Nursing, pharmacy, Healthcare etc. It can be analysis the each documents such as title, abstract, publisher date, etc. The candidate document selection to compare the each and every document and after comparison of given process it can be detail analysis the next stage. If anyone document is missing while comparing the candidate document we can't full fill the next stage the Query Expansion is based on the IR techniques.

More than 5500 biomedical journals are indexed in MEDLINE. New journals are not included automatically or immediately. Selection is based on the recommendations of a panel, the literature selection technical review committee based on scientific scope and quality of a journal[1]. The database contains information such as its name abbreviations and publisher about all journals included in Entrez including pubmed.

Semantic expansion is a technique in information retrieval that adds words similar in meaning i.e. synonyms for better understanding of text. It expands a query with similar words with same meaning. With new added words the information imbibed in the text is easy to reflect the concept and retrieve from huge databases. The widely used text mining approaches are based on similarity of words. However this approach suffers from a serious drawback of vocabulary uses. Terms used by one person to describe a concept may be different when used by another person for describing the same concept.

For example, an author may use the term 'computer display' for monitor while another may use the term 'display screen' for the same. If a document has to be retrieved only on the basis of terms, possibly most of the documents may remain untouched due to lack of term present in it. Though, they may be similar in concept. Semantic search has been identified and recognized as a possible solution for such kind of search where the emphasis is not on text (e.g plagiarism detection) but in finding the concepts and knowledge prevailing in the texts or documents.

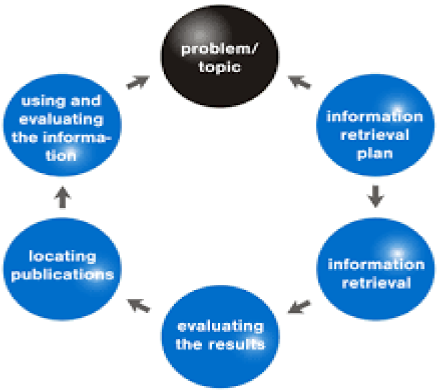

Fig 1: Information Retrieval

WSD task is a potential *intermediate task* for manyother NLP systems, including mono and multilingual Information Retrieval, Information Extraction, Machine Translation or Natural Language Understanding. WSD typically involves two main tasks.

i. Determining the different possible senses (or meanings) of each word.
ii. Tagging each word of a text with its appropriate sense with high accuracy and efficiency.

All methods build a representation of the examples to be tagged using some previous information. The difference between them is the source of this information. The WSD community accepts a classification of these systems in two main general categories:

a) knowledge-based
b) *corpus-based* methods.

In Knowledge-based Method, mainly try to avoid the need of large amounts of training materials required in supervised methods[2][3]. Machine-Readable Dictionaries (MRDs) provide a ready-made source of information about word senses and knowledge about the world, which could be very useful for WSD and NLU.MRDs contain inconsistencies and are created for human use, and not for machine exploitation. There is a lot of knowledge in a dictionary only really useful when performing a complete WSD process on the whole definitions.

Corpus-based Approach, these approaches are those that build a classification model from examples.These methods involve two phases: *learning* and *classification*. The learning phase consists of learning a sense classification model from the training examples. The classification process consists of the application of this model to new examples in order to assign the output senses. Most of the algorithms and techniques to build models from examples come from the Machine Learning area of AI.

One of the first and most important issues to take into account is the representation of the examples by means of features/attributes. That is, which information could and should be provided to the learning component from the examples. The representation of examples highly affects the accuracy of the systems. It seems to be as or more important than the learning method used by the system.

.

II. BACKGROUND

A. Knowledge Sources

Knowledge sources used for WSD are either lexical knowledge released to the public, or world knowledge learned from a training corpus.

1. Lexical Knowledge

In this section, the components of lexical knowledge are discussed. Lexical knowledge is usually released with a dictionary. It is the foundation of unsupervised WSD approaches.

2. Sense Frequency

It is the usage frequency of each sense of a word. Interestingly, the performance of the naïve WSD algorithm, which simply assigns the most frequently used sense to the target, is not very bad. Thus, it often serves as the benchmark for the evaluation of other WSD algorithms.

B. Learned World Knowledge

World knowledge is too complex or trivial to be verbalized completely. So it is a smart strategy to automatically acquire world knowledge from the context of training corpora on demand by machine learning techniques. The frequently used types of contextual features for learning are listed below.

1. Indicative Words

It surround the target and can serve as the indicator of target senses. In general, the closer to the target word, the more indicative to the sense. There are several ways, like fixed-size window, to extract candidate words.

2. Syntactic Features

It refer to sentence structure and sentence constituents. There are roughly two classes of syntactic features. One is the Boolean feature; for example, whether there is a syntactic object. The other is whether a specific word appears in the position of subject, direct object, indirect object, prepositional complement, etc. (Hasting, 1998; Fellbaum, 2001).

3. Domain-specific Knowledge

It is like selectional restrictions, is about the semantic restrictions on the use of each sense of the target word. However, domain-specific knowledge can only be acquired from training corpora, and can only be attached to WSD by empirical methods, rather than by symbolic reasoning. Hasting (1998) illustrates the application of this approach in the domain of terrorism.

4. Parallel Corpora

Parallel corpora is also called bilingual corpora, one serving as primary language, and the other working as a secondary language. Using some third-party software packages, we can align the major words (verb and noun) between two languages. Because the translation process implies that aligned pair words share the same sense or concept, we can use this information to sense the major words in the primary language.

There are no significant distinctions between lexical knowledge and learned world knowledge. If the latter is general enough, it can be released in the form of lexical knowledge for public use. Usually, unsupervised approaches use lexical knowledge only, while supervised approaches employ learned world knowledge for WSD. Examining the literature, however, we found the trend of combination of lexical knowledge and learned world knowledge in recently developed WSD models.

III. PROPOSED APPROACH

This section presents the IR-based approach to the identification of candidate source documents followed by a description of how it can be extended by query expansion using resources from the medical domain.

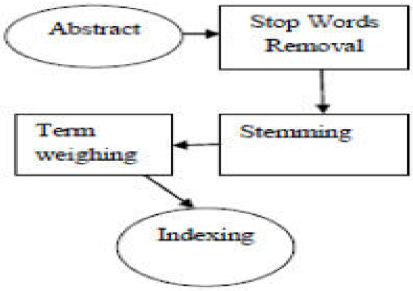

Fig 2. plagiarism detection based on patents matching

A. IR-Based Approach

The process of retrieving candidate source documents using the proposed IR-based approach.The source collection is indexed with an IR system. In the IR-based framework, candidate retrieval process can be divided into four main steps pre-processing, (2) query formulation, (3) retrieval and (4) results merging. These steps are described as follows:

1. Pre-Processing:

Each suspicious document is split into sentence using NLTK. The terms in each sentence are converted to lower case.stopword and punctuation marks are removed.

Stemming Algorithm

A stemming algorithm is a process of linguistic normalisation, in which the variant forms of a word are reduced to a common form, for example,

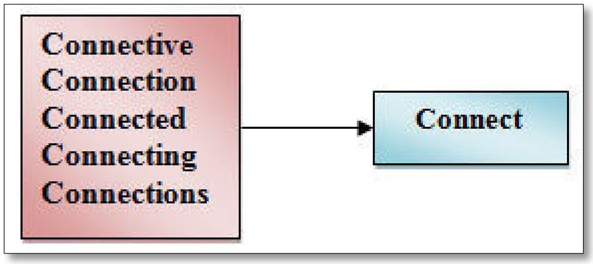

Fig 3. Stem Word

It is important to appreciate that we use stemming with the intention of improving the performance of IR systems. It is not an exercise in etymology or grammar. In fact from an etymological or grammatical viewpoint, a stemming algorithm is liable to make many mistakes. In addition, stemming algorithms - at least the ones presented here - are applicable to the written, not the spoken, form of the language.

2. Query Formulation

Sentences from the suspicious document are used to form multiple queries.The length of a query can vary from a single sentence to all sentences appearing in a document as reused text can be sourced from one or more documents and vary from a single sentence to an entire document . A long query is likely to perform well in situations when large portions of text are reused for plagiarism; on the other hand small portions of plagiarized text are likely to be effectively detected by a short query. Therefore, the choice of query length is important in obtaining effective results.

3. Retrieval:

Terms are weighted using the tf.idf weighting sceme and then text forming the query is used to retrieve similar documents from the index.

4. Result Merging:

The top N documents returned against multiple queries are merged to generate a final ranked list of source documents. A standard data fusion approach, CombSUM, is used to generate the final ranked list of documents by combining the similarity scores of source documents retrieved against multiple queries.

In CombSUM the final similarity score, Sfinalscore, is obtained by adding the similarity scores of source documents obtained against each query q:

$$S_{\text{finalscore}} = \sum_{q=1}^{Nq} Sq(d) \qquad \textbf{(1)}$$

Where,

Nq is the total number of queries to be combined.

Sq(d) is the similarity score of a source document d for a query q.

The top K documents in the ranked list generated by the CombSUM method are marked as potential candidate source documents

B. Query Expansion

The Unified Medical Language System4 (UMLS), a set of tools and resources to assist with the development of biomedical text processing systems, is used to carry out query expansion. Our approach uses two main UMLS resources (the Metathesaurus and MetaMap) which are now described,followed by an explanation of how they are used for queryexpansion.

UMLS Metathesaurus

The UMLS Metathesaurus is a large database of more than 100 multi-lingual controlled source vocabularies and classifications, which contains information about concepts (related to biomedical and health), concept names and relationships between concepts[4][5]. The basic units of the Metathesaurus are concepts, whereby the same concept can be referred to using different terms.

One of the main goals of Metathesaurus is to group all the equivalent terms (synonyms) from different source vocabularies into a single concept. Thus, a concept is a collection of synonymous terms. Each concept in Metathesaurus is assigned a unique identifier called a CUI (Concept Unique Identifier).

1. *Wikitionary*

Wiktionary is a lexical and multilingual dictionary. It was a project of Wikipedia. However unlike Wikipedia which is an encyclopedia, Wiktionary focus on the lexical relation between terms. Wiktionary have certain relationship followed between terms like hyponym, synonyms. Wiktionary is considered as a great source in research associated with ontology and sentiment analysis. For semantic expansion we choose path length approach as previously used in WordNet. Path length method determines the length of path between two nodes which are represented as concepts. Nodes are represented as concepts whereas edge shows the semantic relation between two concepts. It is formulated as in eqn 1,

$$Path\ length\ (con1,con2) = Max.length - Length\ (Con1,Con2) \qquad \textbf{(2)}$$

Where,

Max.length is the longest path in the lexical network of Wiktionary. Length (con1, con2) is the number of edges between paths from concept 1 to concept 2. Wiktionary files are accessed with the help of Java API. It works on available dumps file. We accessed these dump files till March 2015 for our research.

Finding Similarity

The main motive behind semantic expansion is to provide an external source to the query for augmentation. In such situation selection of proper external source becomes very important. Once the query is expanded with external source, next task is to find the similar patents. To find the similarity between concepts we applied Wu-Palmer method.

Wu-Palmer Measure

The principle underlying behind Wu-Palmer model is two concepts can be regarded as similar if they have common source of origination in a taxonomy hierarchy of ontology or lexical dictionary. Formula for Wu-palmer calculation is shown below in eqn 2:

$$Sim\ (c1,c2) = 2H/N1+N2+2H \qquad \textbf{(3)}$$

Where,

N1 and N2 is the no of IS-A relation links from c1 and c2 respectively to the most specific common concept c, and H is the number of IS-A links from c to the root of the taxonomy. It scores between 1 (for similar concepts) and 0. The figure below (2) shows the calculation of Wu-Palmer.

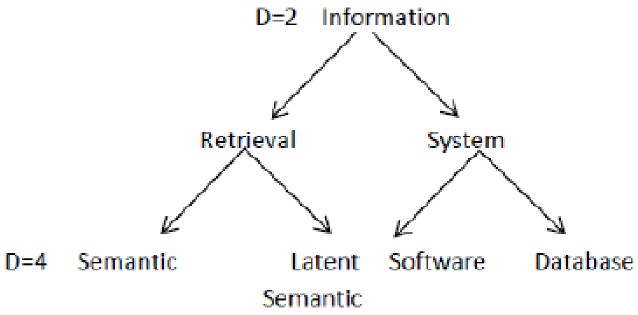

Fig 4. Calculation of Wu-Palmer

Sim (Semantic, Latent Semantic) = 2 * Retrieval/ d (semantic) + d (latent semantic)
=2 x 3/4+4=0.75
Sim (semantic, Software) = 2 * information/ d (software) + d (semantic)
=2 x 2/4+4=0.5

Wu-Palmer not only follows the same idea but also ponders the result with the ratio of the number of term occurrences corresponding to the concepts. In the end, only the documents closest to the query (i.e. whose proximity with the request is higher than a given threshold) are shown to the user, ordered by decreasing similarity. Wupalmer calculation is based on WordNet taxonomy but in our research we applied WordNet as well as Wiktionary as a taxonomy hierarchy.

IV. EXPERIMENTAL SETUP

This section describes the dataset used for evaluation and how the approach was implemented and the evaluation measure used to evaluate the various query expansion methods.

A. Evaluation Dataset

Evaluation is carried out using an existing source of potentially plagiarised publications from Medline. For these experiments, the source collection is fromed from 19,569,568 citations from the 2011MEDLINE/PubMed Baseline Repository. The collection of suspicious documents contains 260 citations from the Deja vu database that have been manually examined and verified as duplicates. These citation pairs are selected because they do not have a common author, making them potential cases of plagiarism.

B. Implementation

Lucene8, a popular and freely available IR system, is used for the experiment. The source collection is indexed. Documents are pre-processed by converting the text into lower case and removing all non-alphanumeric characters. Stopwords9 are removed and stemming is carried out using the Porter Stemmer.Terms are weighted using the tf.idf weighting scheme.Lucene computes the similarity score between query and document vectors.

Lucene computes the similarity score between query and document vectors using the cosine similarity measure:

$$sim(d,q) = \frac{\vec{q} \bullet \vec{d}}{|\vec{q}| \times |\vec{d}|} = \frac{\sum_{i=1}^{n} q_i \times d_i}{\sqrt{\sum_{i=1}^{n}(q_i)^2 \times \sum_{i=1}^{n}(d_i)^2}} \quad (4)$$

where $\overrightarrow{|q|}$ and $\overrightarrow{|d|}$ represent the lengths of the query and document vectors respectively.

C. Evaluation Measure

The goal of the candidate document retrieval task is to identify all the source document(s) for each suspicious document while returning as few non-source documents as possible. It is important for all source documents to be included in the top ranked documents returned by the system since otherwise they will not be identified during later stages of processing [9]. Consequently, recall is more important than precision for this problem. Recall for the top K document, averaged across queries is used as the evaluation measure for these experiments[10]. For a single query the Recall at K (R@K) is 1 if the source document appears in the top K documents retrieved by the query, and 0 otherwise. For a set of N queries, the averaged recall at K score is calculated as:

$$R@K_{avg} = \frac{1}{|N|}\sum_{i=1}^{N} R@K_i \quad (5)$$

where R@Ki is the recall at K score for query i.

V. RESULT ANALYSIS

Our proposed IR-based approach for retrieving candidate documents performs well in identifying real cases of plagiarism. Performance further improves when query expansion is applied.

Table 1 shows no of patents retrieved by the two models individually and third column shows a combined effect of Senticnet and Wiktionary. Query and threshold used in both the tables are same. Combined model of Senticnet and Wiktionary has a better recall rate.

TABLE I
NO. OF RETRIEVED PATENTS FROM THE PATENT DATABASE AFTER SEMANTIC EXPANSION THROUGH SENTICNET, WIKTIONARY AND COMBINED TOGETHER

Query	Q 1	Q 2	Q 3	Q 4	Q 5	Q 6	Q 7	Q 8	Q 9
Senticnet	5	11	5	8	9	12	18	16	10
Wikitionary	19	11	4	10	7	6	12	10	11
Senticnet & Wikit.	11	18	17	19	12	10	8	11	17

Table 2, 3 & 4 shows the effect of the two similarity models on the various external source of expansion. Table 2 shows the similarity models results with Senticnet. Table 3 shows the effect of similarity on Wiktionary and table 4 shows the similarity model effect on combination of Senticnet and Wiktionary expansion. Results are compared in terms of Precision and recall rates. Recall and precision are value of single metrics, so it is better to consider average precision and mean average precision method into consideration which returns a ranked list result, for the accurate measurement of model performance. We also computed the mean recall and mean precision.

TABLE II
SIMILARITY RESULT OF VSM AND WU-PALMER ON SEMANTIC EXPANSION THROUGH SENTICNET

Query	Mean recall	Mean precision	Average precision	Mean Average & Precision
Existing (VSM)	58.1%	40%	4.56%	24.32%
Proposed (Wu-Palmer)	95.6%	70%	5.98%	18.8%

TABLE III
SIMILARITY RESULT OF VSM AND WU-PALMER ON SEMANTIC EXPANSION THROUGH WIKITIONARY

Query	Mean recall	Mean precision	Average precision	Mean Average & Precision
Existing (VSM)	81.2%	40%	4.56%	24.32%
Proposed (Wu-Palmer)	89%	70%	5.98%	18.8%

TABLE IV
SIMILARITY RESULT OF VSM AND WU-PALMER ON SEMANTIC EXPANSION THROUGH SENTICNET & WIKITIONARY

Query	Mean	Mean	Average	Mean

	recall	precision	precision	Average & Precision
Existing (VSM)	84.2%	5.98%	6.66%	24.32%
Proposed (Wu-Palmer)	93%	6.11%	6.69%	18.8%

VI. CONCLUSIONS

In this paper, an attempt was made to identify similar patents through parent and concept vector. This research involved Indian Patent database. Patent abstract is used as search query which is further augmented using external source Senticnet and Wiktionary. This research has two important aspect to focus; first selection of a model for query expansion, secondly selecting the most efficient model which can work in finding similar patents after expansion. We found Wu-Palmer model most efficient for this purpose in comparison with traditional cosine similarity model.

VII. REFERENCES

[1] Rao Muhammad Adeel Nawab,Mark Stevenson,Paul Clough.An IR-based Approach Utilising Query Expansion for Plagiarism Detection in MEDLINE.IEEE/ACM Transaction on Computational Biology and Bioinformatics.(volume pp,Issue:99),March 2016.

[2] G. Amati and C. J. Van Rijsbergen. Probabilistic models of information retrieval based on measuring the divergence from randomness. ACM Transactions on Information Systems (TOIS), 20(4):357–389, October 2002.

[3] N. Balasubramanian, G. Kumaran, and V. R. Carvalho. Exploring reductions for long web queries. In Proc. of SIGIR, pages 571–578, 2010.

[4] M. Bendersky and W. B. Croft. Discovering key concepts in verbose queries. In Proc. of SIGIR, pages 491–498, 2008.

[5] M. Bendersky, D. Fisher, and W. B. Croft. UMass at REC 2010 Web Track: Term Dependence, Spam Filtering and Quality Bias. In Proc. of TREC-10, 2011.

[6] M. Bendersky, D. Metzler, and W. B. Croft. Learning concept importance using a weighted dependence model. In Proc. Of WSDM, pages 31–40, 2010.

[7] M. Bendersky, D. Metzler, and W. B. Croft. Parameterized concept weighting in verbose queries. In Proc. of SIGIR, 2011.

[8] G. Cao, J.-Y. Nie, J. Gao, and S. Robertson. Selecting good expansion terms for pseudo-relevance feedback. In Proc. Of SIGIR, pages 243–250, 2008.

[9] C. L. Clarke, M. Kolla, G. V. Cormack, O. Vechtomova, A. Ashkan, S. B¨uttcher, and I. MacKinnon. Novelty and diversity in information retrieval evaluation. In Proc. Of SIGIR, pages 659–666, 2008.

[10] C. L. A. Clarke, N. Craswell, and I. Soboroff. Overview of the TREC 2009 Web Track. In Proc. of TREC-09, 2010.

[11] W. B. Croft, M. Bendersky, H. Li, and G. Xu. Query representation and understanding workshop report. SIGIR Forum, December 2010.

[12] F. Diaz and D. Metzler. Improving the estimation of relevance models using large external corpora. In Proc. of SIGIR, pages 154–161, 2006.

[13] J. Guo, G. Xu, H. Li, and X. Cheng. A unified and discriminative model for query refinement. In Proc. of SIGIR, pages 379–386, New York, NY, USA, 2008.