IMPROVING PLAGIARISM DETECTION THROUGH THE DOCUMENT CONTENT SEGMENTATION

Tomáš Kučečka, Daniela Chudá

Abstract: Topic extraction from text documents is an important object of research these years. It can be used, for instance, in an information retrieval to group similar results together returned by a web browser or to simplify plagiarism detection in a given document corpus. In this paper we propose a novel approach to document segmentation that should be helpful with plagiarism detection. It is based on analysing keyword positions in a text document. Based on these analyses we try to segment document's content and identify similarity between documents based on the similarity between found segments. Therefore, this approach shows how much about a document can be told from the distribution of its keywords. We think that through this approach we are able to better identify relationships between text documents (between their segments) and come to the more accurate plagiarism identification in a given document corpus. We believe that two documents with similar segments should gain more attention in the plagiarism detection process. In this article we present our first experiments with this approach that we carried out on the student's assignments. We used results from an existing plagiarism detection system PlaDeS to evaluate our approach.

Introduction

Plagiarism is becoming a serious problem these days especially in academic environment. Academic institutions have to take actions to prevent and detect such behaviour in order to keep their credibility. This has become much more difficult with the availability of online resources, which number rapidly increases every year. This makes it a lot easier for people to plagiarise.

Although many people think of plagiarism as rewriting someone else's work without giving the credit to the original author, this is not its only meaning. According to the plagiarism.org resource, the term *"plagiarism"* can be defined as "to steal and pass off (the ideas or words of another) as one's own" or "to use (another's production) without crediting the source".

For the two main problems that are associated with plagiarism detection we consider source identification and paraphrasing. The source identification problem has become much more serious with the spread of Internet and availability of online resources. Without having a source document from which the plagiary comes, it is very difficult to detect a plagiarism. Paraphrasing is another main problem that can be done on different levels. The higher the paraphrasing level, the more difficult is its identification. In this paper we focus on the second problem—paraphrasing identification.

We propose a novel approach to plagiarism detection among texts written in natural language. This approach is based on document segmentation and is mainly intended for longer documents, e.g. bachelor or diploma thesis. The idea of the solution is to model a text document as a mixture of segments, where each segment reflects a topic on different granularity. A document can contain several of these topics with a minimum of one. Through the segments found in documents, we propose document candidates

that should be checked on plagiarism in other way than using standard approaches. For instance, a 3-gram approach can be used to check plagiarism between documents that are less related based on the found segments. On the other hand, documents that contain higher number of similar segments can be checked using more effective comparison techniques with higher computational complexity.

Our approach distinguishes two types of segments—global and local. Global segment covers a global topic of a document, i.e. the topic of a whole document. It holds that every document has exactly one global segment. Local segments represent a subtopic of a document and their number in a document is arbitrary starting from zero. Based on this, every document contains at least one global segment.

Related work

As our approach deals with plagiarism identification based on document topic detection through document segments, we focus here on related work from this area of research.

Finding relationships between text documents based on their topic has gained a significant importance in last years. The existing approaches in this domain vary from different clustering techniques based on frequent-itemsets to probabilistic approaches. Better results in similarity detection can be achieved by single or multi-document segmentation that try to identify where the topic of a text document changes.

The main drawback of the existing approaches is that they frequently consider document as a single topic. This is, in our opinion, correct for shorter texts but becomes a problem for longer documents. A typical example of a text document with several discussed topics is diploma thesis, which is basically structured on the analysis, design, experimental and technical part. For each of these parts usually a different set of keywords is typical. Our approach models a document such way, that it identifies these different parts and their keywords. Basically, one can consider our approach as a single document segmentation method. This is partially true but our approach also extracts keywords that characterize the document as a whole. We call these keywords global keywords.

When designing our own approach, we were inspired by the existing pLSA (probabilistic Latent Semantic Analyses) approach presented by Hofmann, T. (1999) and LDA (Latent Dirichlet Allocation) approach presented by Blei, M.D. et al. (2003). The LDA is basically an enhancement of pLSA. Both of these approaches assign each document with a number of topics, each with different probability based on the words the documents contain. The background behind these approaches is the mixture model, which models the whole document corpus as a distribution of different topics represented by words from a given corpus. Both pLSA and LDA require that a user must know a number of topics to which documents from corpus are going to be assigned. Our approach has no such restriction.

Ambwani, G. et al. (2010) proposed a novel approach to topic segmentation. The authors hold to the idea that the used vocabulary reflects the topic continuity. A text document using this approach is modelled as a graph of RI's (Relevance Interval). RI represents a range of term influence (basically it is a range where the term occurs at

most). Based on the RI's in the constructed graph (the way that the RI's are connected) a document is segmented. Other work on the topic of document segmentation was presented by Sun, B. et al. (2007). Their introduce MI (Mutual information) and WMI (Weighted Mutual Information) indexes for mutual information computation. This approach belongs to the multi-document segmentation. The MI index measures how distinct the two segments are, which is a different approach compared to the standard approaches based on the similarity calculation (e.g. cosine similarity). The WMI index is simply a weighted version of MI computation that should perform better.

Proposed approach

In this section we describe a novel approach to finding relationships between text documents based on their topic. The overall design emerged from our previous work presented in Kueka, T. (2012) and Kueka, T. et al. (2012). We assume only documents written in natural language. Basically, we can divide the whole approach to three main phases:

- Preprocessing
- Segmentation
- Relationships identification

In the preprocessing phase of a document we used standard techniques such as stop words removal, lemmatisation and synonyms replacement. In segmentation phase we identify document segments that represent the document's topics and subtopics. Then, in the third phase we identify relations between segments based on the keywords these segments share.

Segmentation process

The segmentation process is based on calculation of keyword distribution vectors over text blocks, where text block represents a string of a 260 letters. This length was estimated empirically. Text blocks are extracted from document's content. After the extraction process is finished, the following steps are applied:

1. *Keyword extraction.* Based on the tf-idf value a set of keywords is extracted. For every document words with higher weight than 0.03 are selected. Notice, that this threshold is quite low and the number of selected words for a document can therefore be very high. Our aim is not to select as few words as possible but rather to filter out only common words. From now on, we refer to all selected words in this step as keywords.
2. *Finding keyword distribution vector.* The frequencies of keywords (tf values) over extracted text blocks are first calculated. This calculation is done per document. By dividing the keyword frequencies over text blocks (TB) with the total tf value of a keyword in a document, we get a distribution vector. This vector represents a distribution of a keyword over document TBs. In the left picture (a) of Figure 1 an example of a distribution vector of a keyword *w* is shown. The right picture (b) shows two similar distributions of two different words.

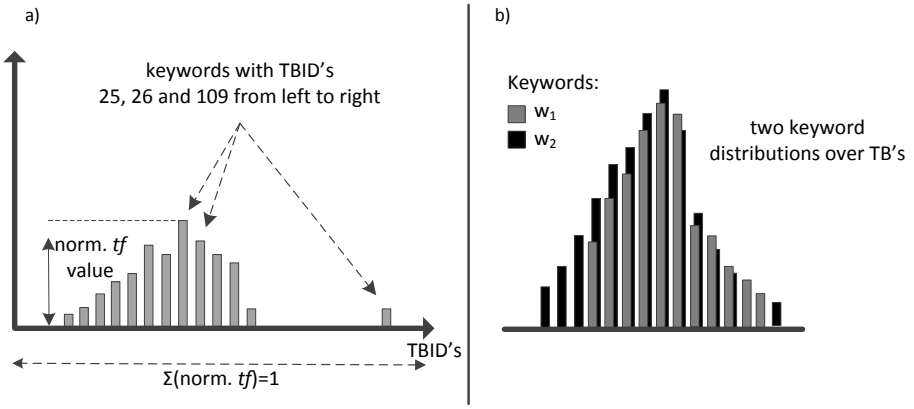

Figure 11.2. a) An example of a keyword distribution; b) Shows two similar distributions of two different keywords

By TBID we mean the ID of a text block. Every text block has an ID that is unique per document. This ID represents the order of a text block in a document. The first extracted TB has ID 0.

3. *Comparison of keyword distributions.* We compare the keyword distributions per document using the Jensen-Shannon (JS) divergence presented in Cedeño, B.A. et al. (2009) to find out if two keywords occur in a text document in similar text blocks. To such keywords we refer as similar. The output of JS distance is a value that represents the amount of work that has to be done to transform one distribution vector on another. This value lies in the [0, 1] interval.
4. *Segment building.* Keywords with similar distribution form a segment, i.e. every segment is represented by a set of similar keywords. In order to tell if a keyword belongs to a local or a global segment, we calculate the variance for all distribution vectors. Higher variance indicates that the keyword is more sparsely present in a document's content. In such case we consider the keyword global. The distributions with highest variance form a global segment of a document. Other keywords are local and therefore can form local segments. Two different distributions of a keyword, one of which is local and one of which is global, are depicted in Figure 2.

A minimum number of keywords in local segment is 3. This value was estimated empirically based on the experiment results. Otherwise the segment will not be created. All local keywords that do not belong to any segment are dropped.

Relationships identification

First, an undirected segment graph G is constructed from found segments. Vertices of G represent segments and edges relations between these segments. A relation is a triple (x, y, z) where x, y are similar segments and z represents their similarity. Based on the segment types, we distinguish three types of relations in this graph:

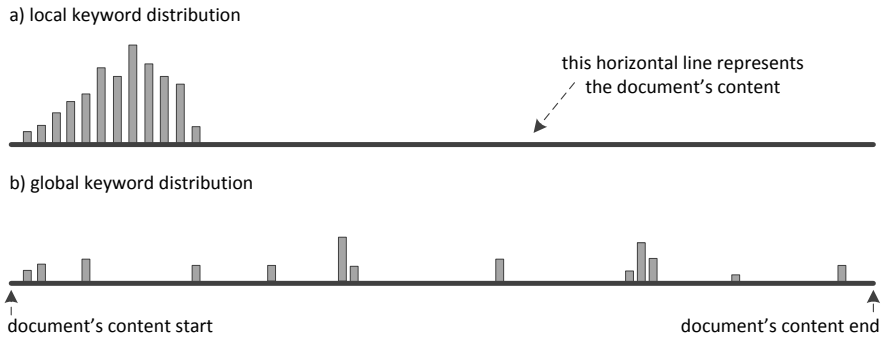

Figure 11.2. An example of two keyword distributions where a) is local and b) is global. It is clear that the distribution variance is higher in case of b). Therefore in this case we consider the keyword global

- *Parent relations*—exists between similar local and global segments coming from different documents (*x* is local segment, *y* global or vice versa).
- *Global relations*—exists between similar global segments (both *x* and *y* are global).
- *Local relations*—exists between similar local segments (both *x* and *y* are local).

Two segments are related if they share some minimum number of keywords. This value can differ for different types of relations. For now, the minimum similarity is 0.25 and was set based on the results of carried experiments.

Found relations are stored in a segment graph in a form of an edge connecting two segments. This edge is associated with a value ranging [0, 1] which represents a relation strength. If two segments are related, they exchange their keywords. For instance, segment *S*1 contains *keywords goal, question, metric, GQM, success* and segment *S*2 *goal, question, metric, GQM, project*. These two segments are related, therefore, they exchange their keywords. The result after the exchange will be *S*1 *goal, question, metric, GQM, success, project* and *S*2 *goal, question, metric, GQM, project, success*. The whole situation is depicted in the Figure 3.

Experiments

This section describes the carried experiments with the proposed model. First we explain how we determined model parameters, e.g. text block length, tf-idf threshold. Then we show how well our approach performed when we compared its results to the output of plagiarism detection system PlaDeS described by Chudá, D. et al. (2011). PlaDeS is a plagiarism detection system developed at Slovak University of Technology, Faculty of Informatics and Information Technologies. We decided to use this system because it performs plagiarism check over documents written in Slovak language and because we know details about this system as we are its authors.

We had two datasets on which we carried out our experiments. The first dataset contains articles from web written in English language, the second one consists of student assignments written in Slovak language.

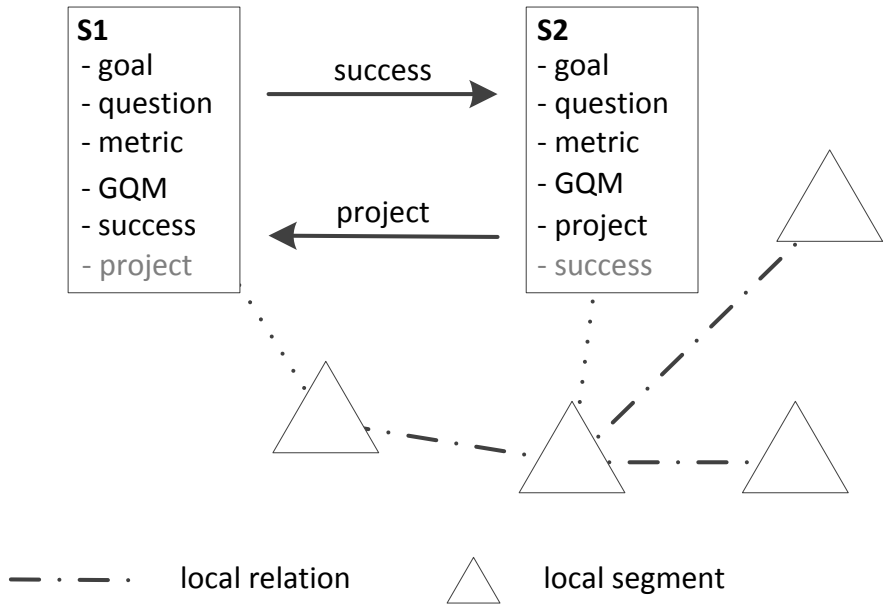

Figure 11.2. An example of a segment keyword exchange in the created segment graph. The keywords with lighter colour in the tables are those that were exchanged

Model parameter estimation

For the parameter estimation of the proposed model we used web articles from the BBC Travel. We selected 220 web pages, each web page dealing with exactly one topic, and manually annotated them. The annotated articles were about 600 words long on average. The output of the annotation process was a set of at least 5 to max 10 keywords associated with every web article.

Here is the list of the parameters that we estimated empirically on this dataset. The number after the hyphen represents the estimated parameter value for which we achieved the best results.

- text block (TB) length in letters—260
- tf-idf threshold—0.03
- minimum number of keywords in a local segment—3
- maximum number of keyword in global segment—6
- minimum threshold for local-to-local segment similarity in order to be related—0.3
- minimum threshold for global-to-global segment similarity in order to be related—0.3

To estimate these parameter values we used our approach to find all global keywords for every web article and compared the output with the user's keywords (user's

keywords are those, that were the output of the annotation process). As a global keyword we considered top 5 keywords with the highest variance returned by our method. The tf-idf approach achieved precision 0.40 and recall 0.27. Our approach achieved precision 0.46 and recall 0.30. Clearly our approach outperformed the tf-idf in the keyword extraction process.

Performance

To evaluate the overall performance of our approach, we compared its output to the output of a plagiarism detection system PlaDeS. As a detection method in this system we used 3-grams.

PlaDeS returns similar pairs of documents as triples $(d1, d2, s)$, where $d1$ and $d2$ are the compared documents and s represents their similarity. The returned similarity can be different considering the order in which the documents are compared. For instance, the s value for two triples $(d1, d2, s)$ and $(d2, d1, s)$ will differ in most cases (asymmetric similarity). However, our approach that we proposed in this paper does not distinguish the order of the compared documents. Remember that the relation is defined as a triple (x, y, z), where both (x, y, z) and (y, x, z) return the same value z (symmetric similarity). Therefore, we compare the output of our solution with PlaDeS in such way, that if our approach identifies relation (x, y, z), we check if the output of PlaDeS does not contain a pair (x, y, s) or a pair (y, x, s). We do not search for a match in all document pairs retuned by PlaDeS, but only those with similarity higher than 4%. This threshold was estimated based on our own experience with plagiarism checking.

Our dataset contained 313 student assignments written in Slovak language. An average length of each assignment was about 2500 words. The overall performance of our method is depicted in Figure 4. This figure shows the portion of detected suspicious document pairs (y axis) for different similarity thresholds (x axis) when compared to the output of PlaDeS. For instance, if we consider all document pairs returned by PlaDeS with similarity 20% (x axis), our approach detected 60% (y axis) of these pairs. The total number of unordered pairs of documents returned by PlaDeS was 48 828 while our approach detected 14,530 pairs. Figure 4 represents the achieved recall of our system when compared to PlaDeS. However, we were not able to evaluate the precision of our system with this experiment. This is because we cannot tell which document pairs returned by out method belong to the similarity pairs returned by PlaDeS (there is a different similarity calculation used in these two systems).

This high number (14,530) of document pairs detected by our approach is due to the large amount of identical topics. The used dataset consists of student assignments from last 8 years, while in each year there were around 15 different topics. The total number of students during these 8 years was 313. This means, that there can be found many possible similar pairs between students' assignments within different years. This could be one explanation why the carried experiments showed such results (Figure 4).

Conclusion

In this paper we presented a novel approach to plagiarism detection between documents written in natural language. This approach segments documents on subtopics by

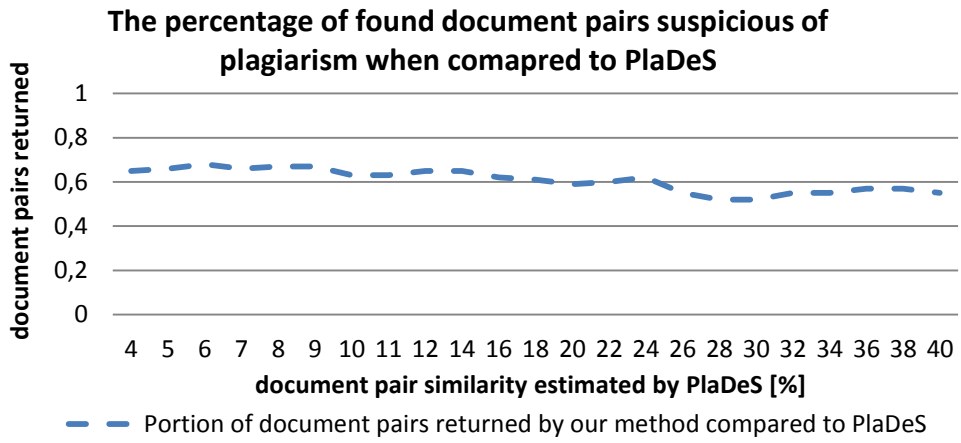

Figure 11.2. Recall of our method. Shows portion of detected suspicious document pairs by our method when compared to the output of a plagiarism detection system PlaDeS

watching keyword distributions in document's content. From the identified segments we construct a segment graph that stores found relationships between these segments belonging to different documents. Based on these relations, document pairs with similar topics are extracted and recommended on special plagiarism checking. By special checking we mean computationally more complex comparison techniques than standard approaches.

The results showed that our approach detected about 60% of all suspicious document pairs returned by a plagiarism detection system PlaDeS. The total number of document pairs that our solution returned is 14 530 which represents the 29.76% of all possible pairs that have to be compared by PlaDeS. If we now wanted to use a more sophisticated method to compare some of the similar segments found in the documents, we would need less computation power than in the case of using PlaDeS

As a main advantage of our approach we consider the characteristics of the proposed algorithm that we use to find relations between segments belonging to different documents. Because we distinguish two types of segments, global and local, the relations between these segments also differ. Probably the biggest benefit this has is that it enables us to deeply explore the found similarities between documents. For instance, we can differently focus on analysing the local-to-global and global-to-global relations in the builded segment graph. Possible drawbacks of the proposed solution are that it is mainly suited for longer text documents. Although its performance on shorter texts has not been determined yet, we expect its drop. Also we have not yet determined what computation cost our solution has in case when comparing thousands of text documents.

In order to improve the plagiarism detection, in the nearest future we decided to integrate the described approach into the plagiarism detection system PlaDeS. We expect from this step to be able to better detect paraphrasing, because more complex approaches for plagiarism detection can be applied on related document segments.

For now, we did not take any advantage in the carried experiments from the different types of relations stored in the graph (local-to-local, global-to-global, etc.). In the future we would also like to use different analyses techniques with different types of relations. For instance, based on the type of found relations in segment graph we might be able to determine if students copy more local areas of documents or tend to copy documents as a whole.

References

Ambwani, G., Davis, AR. (2010) Contextually-mediated semantic similarity graphs for topic segmentation. 2010. *Workshop on Graph-based Methods for Natural Language Processing*, 60–68 2010, Stroudsburg USA.

Blei, MD., Ng, YA., Jordan, IM. (2003) Latent Dirichlet Allocation. *The Journal of Machine Learning Research,* 3, 993–1022.

Cedeño, BA., Rosso, P., Benedí, MJ. (2009) Reducing the Plagiarism Detection Search Space on the Basis of the Kullback-Leibler Distance. *10th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing '09)*, 523–534 2009, Berlin Germany.

Chudá, D., Kučečka, T. (2011) PlaDes—Effective Way to Plagiarism Detection of Student Assignments Written in Natural Language. *1st International Conference on E-learning and the Knowledge Society (E-learning'11)*, 47–52 2011, Bucharest Romania.

Hofmann, T. (1999) Probabilistic latent semantic indexing. *22nd annual international ACM SIGIR conference on Research and development in information retrieval*, 50–57 1999, New York, USA.

Kučečka, T. (2012) Identifikácia previazania slov prostredníctvom ich rozloženia v dokumente [In English: Identifying Keyword Relations Based on Their Position in Text]. *7th Workshop on Intelligent and Knowledge Oriented Technologies*, 45–48 2012, Smolenice Slovakia.

Kučečka, T., Chudá, D. (2012) Identifikácia témy a podtémy dokumentu na základe pozície kľúčových slov v texte [In English: Topic Identification Based on Keyword Positions]. *Annual Database Conference (DATAKON)*, 127–136 2012, Czech Republic.

Sun, B., Mitra, P., Giles, CL., Yen, J., Zha, H. (2007) Topic segmentation with shared topic detection and alignment of multiple documents. *30th annual international ACM SIGIR conference on Research and development in information retrieval*, 199–206 2007, New York, USA.

Acknowledgements

This work was partially supported by the Scientific Grant Agency of the Slovak Republic, grant No. VG1/0971/11 and is the partial result of the Slovak Research and Development Agency under the contract No. APVV-0208-10.

Authors

Tomáš Kučečka, kucecka@fiit.stuba.sk,
Daniela Chudá, chuda@fiit.stuba.sk
Slovak University of Technology, Slovakia