International Journal of Database Theory and Application
Vol.9, No.12 (2016), pp.35-44
http://dx.doi.org/10.14257/ijdta.2016.9.12.04

A Method of Plagiarism Source Retrieval and Text Alignment Based on Relevance Ranking Model

Leilei Kong[1,2], Zicheng Zhao[2], Zhimao Lu[1], Haoliang Qi[2*] and Feng Zhao[1]

[1]*Colledge of Information and Communication Engineering, Harbin Engineering University*
[2]*Computer Science and Technology, Heilongjiang Institute of Technology*
Email: kongleilei1979@gmail.com

Abstract

The problem of text plagiarism has increased because of the digital resources available on the World Wide Web. Source Retrieval and Text Alignment are two core tasks of plagiarism detection. A plagiarism source retrieval and text alignment system based on relevance ranking model is described in this paper. Not only the source retrieval task but also the text alignment task is all regarded as a process of information retrieval, and the relevance ranking is used to search the plagiarism sources and obtain the candidate plagiarism seeds. For source retrieval, BM25 model is used, while for text alignment, Vector Space Model is exploited. Furthermore, a plagiarism detection system named HawkEyes is developed based on the proposed methods and some demonstrations of HawkEyes are given.

Keywords*: Plagiarism Detection, Source Retrieval, Text Alignment, Relevance Ranking Model*

1. Introduction

The problem of plagiarism has increased because of the digital resources available on the World Wide Web [1]. During the last decade, research on automated plagiarism detection in natural languages has actively evolved, which takes the advantage of recent developments in related fields like information retrieval (IR), cross-language information retrieval, natural language processing, *et al*. At the same time, the increasingly serious problem of plagiarism accelerated the development of plagiarism detection softwares.

Particularly remarkable attention about this field is the plagiarism detection algorithms evaluation organized by Cross Language Evaluation Forum(CLEF) speeds up the development of plagiarism detection algorithms and the related works. Plagiarism detection, with author identification and author profiling, become known as PAN(International Evaluation Competition on Uncovering Plagiarism, Authorship, and Social Software Misuse) in CLEF[1].

PAN proposed a general framework of plagiarism detection [2-3]. In this framework, most plagiarism detection algorithms contain two main tasks: source retrieval and text alignment. Given a suspicious document and a web search API, the task of source retrieval is to retrieve all plagiarized sources while minimizing retrieval costs. Since we do not know which segments in suspicious document have plagiarized from the source document, we use the words or phrases generated from suspicious document as queries to submit the search engine for retrieving the plagiarism sources. Given a solid search engine, query generation is a core problem of source retrieval. These queries extracted from the suspicious document will be submitted to the search engine for retrieving the

* corresponding author
[1] http://pan.webis.de/

ISSN: 2005-4270 IJDTA
Copyright ⓒ 2016 SERSC

plagiarism sources. These retrieved documents are called candidate source documents. The qualities of the candidate source documents are mainly decided by the qualities of queries mainly.

After searching some candidate plagiarism source documents, given a pair of documents(a suspicious document and a source document), the task of text alignment is to identify all contiguous maximal-length passages of reused text between them. Normally, the algorithms of text alignment first search some matching text segments, called plagiarism seeds. Then the seeds will be merged to get the aligned text segments. In this task, searching plagiarism seeds is one of the most important task.

In this paper, by using the view of information retrieval, we look upon the processes of query generation in source retrieval and searching the plagiarism seeds in text alignment as the problem of computing the relevance between query and document. The queries for source retrieval is generated by using BM25 model and the matching of pair of plagiarism seeds is acquired by exploiting Vector Space Model (VSM).

This rest of the paper is organized as followed. In Section 2, we introduce the related work of source retrieval and text alignment in plagiarism detection. In Section 3, we propose the methods of query generation based on BM25 and the plagiarism seeds acquirement based on VSM. In Section 4, we give a demonstration of a source retrieval and text alignment system developed based on proposed method. In Section 5, we give a conclusion.

2. Related Work

There are three research areas related to our research work: plagiarism detection, source retrieval and text alignment. In this section, we briefly describe the related works on these areas.

2.1. Plagiarism Detection

Figure 1 describe the general retrieval process of plagiarism detection proposed by [2-3]:

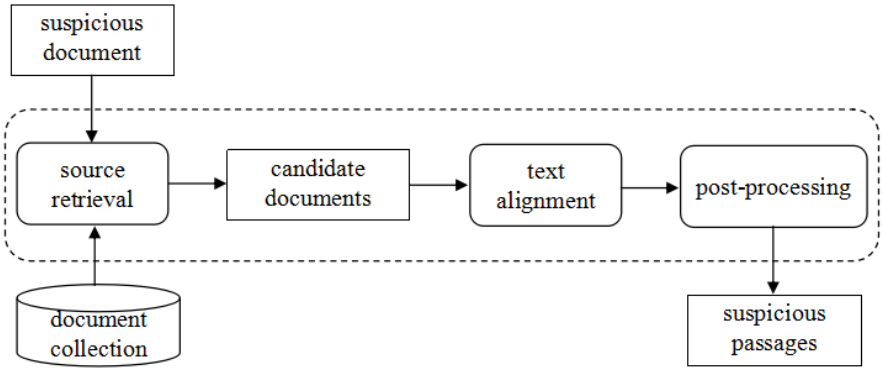


Figure 1. General Retrieval Process of Plagiarism Detection

Given suspicious document d_{susp} and a (very large) document collection D of potential plagiarism source documents, the task of plagiarism detection is to detect by searching for text passages in D that are highly similar to text passages in d_{susp}. In plagiarism detection, there are three basic steps:(1)source retrieval is used to identify a small set of candidate source documents $D_{src} \subseteq D$ that are likely sources for plagiarism regarding d_{susp}. (2) text alignment use the document in D_{src} to compare with d_{susp} for extracting all high similar passages of text. (3) post-processing is used to filter and clean the extracted passage pairs,

 Copyright ⓒ 2016 SERSC

and possibly visualized for later presentation. In this paper, we focus on source retrieval and text alignment.

2.2. Source Retrieval

The process of Source retrieval can be described as Figure 2.

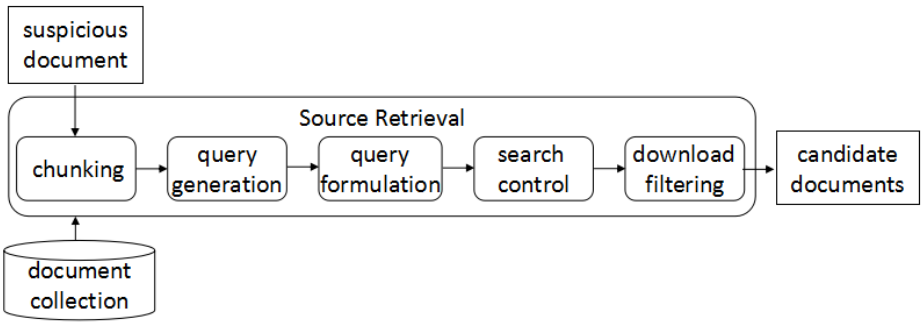


Figure 2. Generic Source Retrieval Process of Source Retrieval

Source retrieval follows the five steps outlined in Figure 2 [2-**Error! Bookmark not defined.**]. (1) *chunking*. Given a suspicious document d_{susp}, chunking divides d_{susp} into smaller passages. (2) *query generation*. Given a passage, keyphrases are extracted from it in order to construct the queries to retrieve plagiarism sources. (3) *query formulation. G*iven sets of keyphrases extracted from passages, queries are formulated in terms of being accepted by the API of the search engine. (4) *search control*. Given a set of queries, the search controller schedules their submission to the search engine and directs the download of search results. (5) *download filtering*. By using a filtering algorithm, download filtering decides which downloaded documents will be further compared in detail with the suspicious document.

In the above source retrieval process, keyphrase extraction is regraded as the most important one for the source retrieval algorithm since the decisions made here directly affect the overall performance [3]. Generally, TFIDF [4], the name entities [5], the terms with rarest frequency on document level [6], or only nouns, adjectives and verbs [7, 8] are used to generate queries.

2.3. Text Alignment

Text alignment can be depicted in Figure 3.

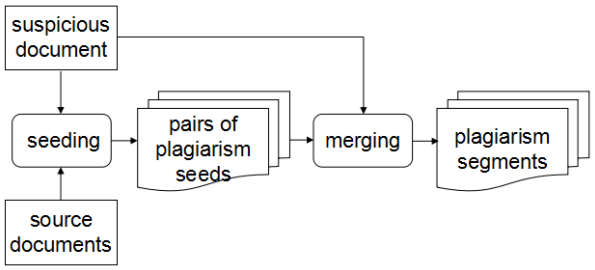


Figure 3. Generic Process of Text Alignment

In Figure 3, seeding is used to search pairs of matches (also called “seeds”) between suspicious document and source documents using some seed heuristic. Then they will be merged into aligned text passages of maximal length by a merging algorithm [2-3]. A number of seed heuristics have been applied. The most frequently used method is to

Copyright ⓒ 2016 SERSC

compare the common lexical features of two documents. For example, Suchomel *et al.* [9] use sorted word 5-grams and unsorted stop word 8-grams. Grozea and Popescu [10] use char 16-grams. Rodríguez Torrejón [11] use sorted word 3-grams and sorted word 1-skip-3-grams. Palkovskii and Belov [12] use word 3-grams.

3. Method of Plagiarism Source Retrieval and Text Alignment Based on Information Retrieval Model

3.1. Relevance Ranking Model

In information retrieval, the goal of a relevance ranking model is to produce a ranked list of documents according to the relevance between these documents and the query [13]. It can be described in Figure 4.

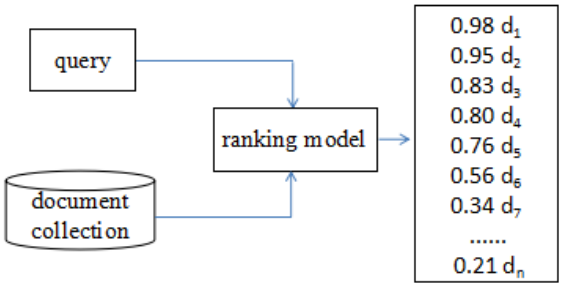


Figure 4. Ranking Document According to the Relevance Ranking Model

The relevance ranking model usually takes each individual document as an input, and computes a score measuring the matching between the document and the query. Then all the documents are sorted in descending order of their scores. In other word, the relevance ranking models is used to measure the relevance of document and query. The higher the similarity score between documents and query, the more relevant they are.

The relevance ranking models retrieve documents based on the occurrences of the query terms in the documents(such as Boolean model), on the relevance degree(such as Vector Space Model), or based on the probabilistic ranking principle(such as BM25 model, or the Language Model).

3.2. Source Retrieval Based on BM25

As described in Section 1 and Section 2, given a search engine, the query generation is the core issue of source retrieval. Inspired by information retrieval model, we regard the word, phrase or word combinations in suspicious segment as the query in information retrieval, and the suspicious segment as the document in information retrieval. Then, the target of query generation is to find out the words which best represents the suspicious segment.

We compare many methods of query generation, such as BM25, Language Model, or only nouns, verbs and adjectives, to find out a better query generation method. We tried constructing queries by ranking keywords by TF, TF-IDF, LM, and BM25 and then combining the top k keywords as queries. We empirically found that BM25 keyword ranking performed the best.

Our approach uses an unsupervised ranking method to rank the words extracted by a query generation method by their similarity to the suspicious document segments. The terms of a suspicious document segment are ranked according to the BM25 methods, and the top n terms are selected as the queries. Then we combine 10 non-overlapping top n terms to generate a query for a suspicious segment. Algorithm 1 shows our source retrieval algorithm.

 Copyright ⓒ 2016 SERSC

Algorithm 1 Source retrieval strategy based on BM25

Input: a set of suspicious document D_{susp}
Output: the results of source retrieval D_{src} for each suspicious document

```
1: for all d_susp ∈ D_susp do
2:   p_susp ← SPLITd_susp (d_susp)
3:   for all p ∈ p_susp do
4:     p ← PREPROCESS(p)
5:     for all words k∈ p do
6:       score(k_i)← BM25(k,p)
7:       wordsList ← ADD(score_k)
8:     end for
9:     wordsList ← RANK(wordsList)
10:     for i = 1..topn do
11:       query_p ← ADDWORDS(wordList(i))
12:   end for
13:  end for
14:   for all query_p do
15:     results ←SUBMITQUERIES(queries[i])
16:    end for
17:     results ← MERGE(results)
18:    for all results do
19:     source ← DOWNLOAD(result)
20:    end for
21: end for
```

In Step 6 of Algorithm 1, we use BM25 to compute the score of each word in some suspicious segment. The basic idea of BM25 is to rank documents by the log-odds of their relevance. BM25 extracts the keyphrases according to the outputs from the Okapi BM25 [14]. Given a term k in suspicious document segment s_k, the BM25 score of term k_i is computed as:

$$score(k_i) = \frac{\mathrm{tf}(k_i, s_k) \cdot \mathrm{idf}(k_i) \cdot (k_1 + 1)}{\mathrm{tf}(k_i, s_k) + k_1 \cdot (1 - b + b \cdot \frac{LEN(s_k)}{avsl})} \quad (1)$$

where $tf(k,s_k)$ is the term frequency of the term k_i in the suspicious document segment s_k, $LEN(s_k)$ is the length (number of words) of suspicious document segment s_k, and avsl is the average segment length in the suspicious documents segments collection. k_1 and b are free parameters and we set k_1=1.2, b=0.75. $idf(k_i)$ is the IDF weight of the term k_i, computed by using (9).

$$IDF(k_i) = \log(\frac{N}{n(k_i)}) + 1 \quad (2)$$

where N is the total number of documents in the corpus C, and $n(k_i)$ is the number of documents containing the term k_i.

The terms of a suspicious segment are ranked according to the above BM25 method, and queries are constructed by combining each non-overlapping top n terms, here we set n=10. Then the top 10 most similar words with suspicious segment are chosen and combined into a query. They will be submitted to the search engine to retrieve the source documents.

3.3. Text Alignment Based on VSM

Copyright ⓒ 2016 SERSC

Seed searching identifies exact smaller matches between suspicious document and its source document [2-3]. Followed the classical methods described in Section 2, we firstly try to obtain the plagiarism seeds as precise as possible. We choose the sentences as the base plagiarism comparing units. We view the sentence in suspicious document as the query and the sentences in source document as the documents. Then the problem of identifying the plagiarism seeds is formalized as a problem of retrieving the most relevant documents (the sentences in source document) when given a query(the sentence in suspicious document), which make us use the retrieval model to achieve our goals. Then the problem of obtaining the pairs of plagiarism seeds is converted into ranking the sentences in source document according to the relevance of between the documents and the query. Then the top n most relevant sentences will be chosen as the candidate plagiarism seeds.

In our method, the Vector Space Model (VSM) is used as the retrieval model to rank the sentences in source document. In Vector Space Model, both documents and queries are represented as vectors in a Euclidean space, in which the inner product of two vectors can be used to measure their similarities. To get an effective vector representation of the query and the documents, TF-IDF weighting has been widely used. The TF of a term t in a vector is defined as the normalized number of its occurrences in the document, and the IDF is defined in Eq. (1). The VSM uses the cosine distance to measure the relevance of documents and query. The VSM is shown in Eq. (3):

$$Cos(I_s, I_R) = \frac{\sum_{k=1}^{n} w_{S_k} * w_{R_k}}{\sqrt{(\sum_{k=1}^{n} {w_{S_k}}^2)(\sum_{k=1}^{n} {w_{R_k}}^2)}} \quad (3)$$

where I_S and I_R are a pair of sentences from the suspicious document S and the source document R, and and are the weights of terms in S and R receptively. We compare the sentence s_{suspi} in suspicious document with all the sentences s_{srcj} in source document by using VSM to compute the similarity score for a pair of sentence(s_{suspi}, s_{srcj}). Each sentence in source document for which the similarity score computing by VSM was above a threshold t_1 is retained in their position order in source document.

Note that in paraphrasing plagiarism, the content of suspicious documents is obfuscated on purpose. It increases the degree of difficulty to identify the plagiarism seeds. So we choose the technology of information retrieval integrating the synonym recognition. For each noun, we expand it by using HIT-CIR Tongyici Cilin (Extended) if they are written in Chinese[2] and using WordNet if they are written in English, while the sentences in source document are viewed as short documents.

Then, given seed matches identified between a suspicious document and a source document, they will be merged by a text merging algorithm to form a longer candidate plagiarism paragraph. We use Bilateral Alternating Sorting algorithm to merge the candidate plagiarism seeds which described in [15] in detail. Lastly, using a segment filtering algorithm, we remove all the segments which are lower than a threshold t_2 computed by Jaccard coefficient.

Algorithm 2 shows the algorithm of text alignment.

Algorithm 2 Text alignment strategy based on VSM

Input: a suspicious document d_{susp} and a source document d_{src}
Output: the plagiarism segments $plgseg_{susp}$ for d_{susp}

1: for d_{susp} do
2: $s_{susp} \leftarrow SPLITd_{susp}$ (d_{susp})
3: $s_{susp} \leftarrow EXPAND(s_{susp})$
4: end for

[2] http://ir.hit.edu.cn/demo/ltp/Sharing_Plan.htm

 Copyright ⓒ 2016 SERSC

5: for d_{src} do
6: $s_{src} \leftarrow SPLITd_{src}(d_{src})$
7: $s_{src} \leftarrow EXPAND(s_{src})$
8: end for
9: for all s_{suspi} do
10: for all s_{srcj} do
11: if VSM (s_{suspi}, s_{srcj}) > t_1 then
12: $seeds_{susp} \leftarrow (s_{suspi}, s_{srcj})$
13: end if
14: end for
15: end for
16: for all $seeds_{susp}$ do
17: BiRANK ($seeds_{susp}$)
18: end for
19: for all seeds $\in$ $seeds_{susp}$ do
20: $seg_{susp} \leftarrow$ MERGE(seeds)
21: if Jaccard(seg_{susp} > t_2) then
22: $plgseg_{susp} \leftarrow seg_{susp}$
23: end if
24: end for

4. Demonstration of Plagiarism Detection System

Using the proposed method, we develop a plagiarism detection system named HawkEyes. The core functions of HawkEyes include: plagiarism source retrieval, plagiarism text alignment, documents management, detection results visualization, *etc*.

The detailed work flows of HawkEyes are described in Figure 5.

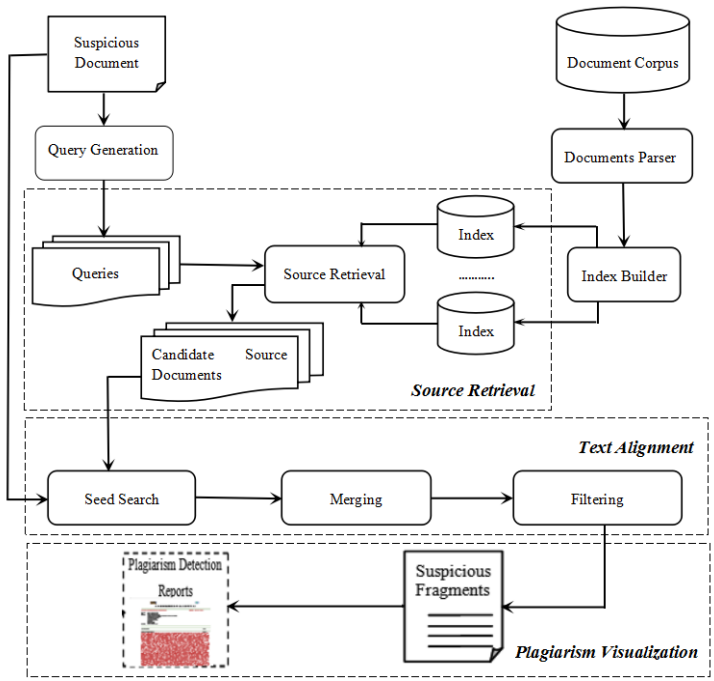
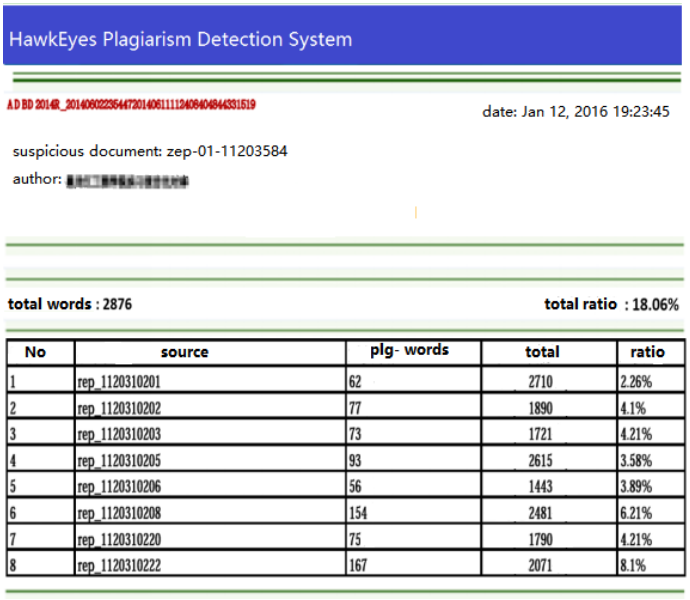


Figure 5. Detailed Work Flow of HawkEyes

Figure 6 gives an example of source retrieval of Chinese plagiarism detection.

Copyright ⓒ 2016 SERSC


HawkEyes Plagiarism Detection System

date: Jan 12, 2016 19:23:45

suspicious document: zep-01-11203584

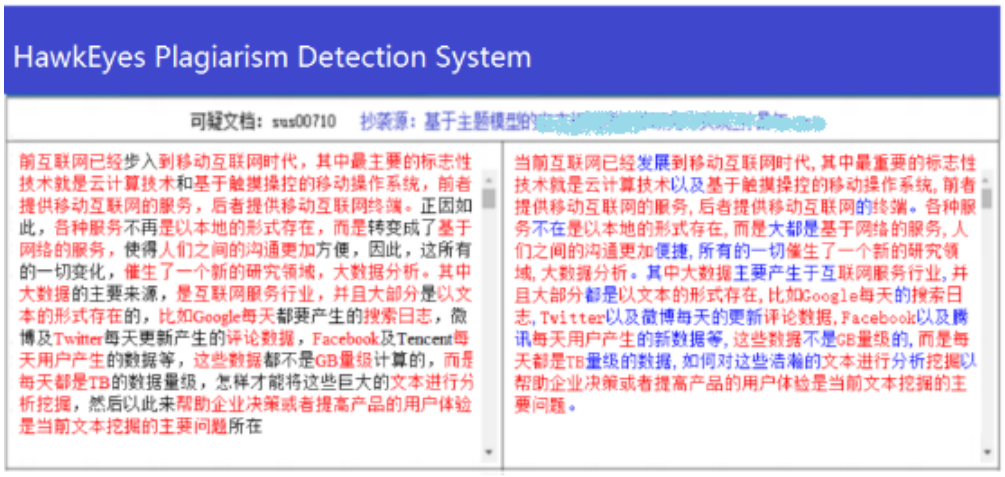
author:

total words : 2876 total ratio : 18.06%

No	source	plg- words	total	ratio
1	rep_1120310201	62	2710	2.26%
2	rep_1120310202	77	1890	4.1%
3	rep_1120310203	73	1721	4.21%
4	rep_1120310205	93	2615	3.58%
5	rep_1120310206	56	1443	3.89%
6	rep_1120310208	154	2481	6.21%
7	rep_1120310220	75	1790	4.21%
8	rep_1120310222	167	2071	8.1%

Figure 6. Source Retrieval Results Demonstration

And Figure 7 shows a text alignment results between a suspicious document and its plagiarism source document.


HawkEyes Plagiarism Detection System

可疑文档：sus00710 抄袭源：基于主题模型的

前互联网已经步入到移动互联网时代，其中最主要的标志性技术就是云计算技术和基于触摸操控的移动操作系统，前者提供移动互联网的服务，后者提供移动互联网终端。正因如此，各种服务不再是以本地的形式存在，而是转变成了基于网络的服务，使得人们之间的沟通更加方便，因此，这所有的一切变化，催生了一个新的研究领域，大数据分析。其中大数据的主要来源，是互联网服务行业，并且大部分是以文本的形式存在的，比如Google每天都要产生的搜索日志，微博及Twitter每天更新产生的评论数据，Facebook及Tencent每天用户产生的数据等，这些数据都不是GB量级计算的，而是每天都是TB的数据量级，怎样才能将这些巨大的文本进行分析挖掘，然后以此来帮助企业决策或者提高产品的用户体验是当前文本挖掘的主要问题所在

当前互联网已经发展到移动互联网时代，其中最重要的标志性技术就是云计算技术以及基于触摸操控的移动操作系统，前者提供移动互联网的服务，后者提供移动互联网的终端。各种服务不在是以本地的形式存在，而是大都是基于网络的服务，人们之间的沟通更加便捷，所有的一切催生了一个新的研究领域，大数据分析。其中大数据主要产生于互联网服务行业，并且大部分都是以文本的形式存在，比如Google每天的搜索日志，Twitter以及微博每天的更新评论数据，Facebook以及腾讯每天用户产生的新数据等，这些数据不是GB量级的，而是每天都是TB量级的数据，如何对这些浩瀚的文本进行分析挖掘以帮助企业决策或者提高产品的用户体验是当前文本挖掘的主要问题。

Figure 7. Text Alignment Demonstration

5. Conclusion

In this paper, inspired by relevance ranking model, we propose a method for source retrieval and text alignment in plagiarism detection. We formalize the problem of query generation as a ranking words in suspicious segment, while the plagiarism seeding acquirement as a measure the relevance between the sentence in suspicious document and the sentences in source document. Using relevance ranking model BM25 and Vector Space Model respectively, we give the method of source retrieval based on BM25 and the method of text alignment based on VSM. And a plagiarism detection system is realized by using the proposed method.

Copyright ⓒ 2016 SERSC

Acknowledgment

This work is supported by Youth National Social Science Fund of China (No.14CTQ032).

References

[1] S. M. Alzahrani, N. Salim and A. Abraham, "Understanding plagiarism linguistic patterns, textual features, and detection methods", IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, vol. 42, no. 2, **(2012)**, pp. 133-149.
[2] M. Potthast, T. Gollub and M. Hagen, "Overview of the 4th International Competition on Plagiarism Detection", CLEF (Online Working Notes/Labs/Workshop), **(2012)**.
[3] M. Potthast, M. Hagen and T. Gollub, "Overview of the 5th international competition on plagiarism detection", CLEF Conference on Multilingual and Multimodal Information Access Evaluation. CELCT, **(2013)**, pp. 301-331.
[4] L. Kong, H. Qi, S. Wang, C. Du, S. Wang and Y. Han, "Approaches for Candidate Document Retrieval and Detailed Comparison of Plagiarism Detection Notebook for PAN at CLEF 2012", CLEF 2012 Evaluation Labs and Workshop – Working Notes Papers, Rome, Italy, **(2012)**.
[5] V. Elizalde, "Using Statistic and Semantic Analysis to Detect Plagiarism—Notebook for PAN at CLEF 2013", CLEF 2013 Evaluation Labs and Workshop – Working Notes Papers, Valencia, Spain, **(2013)**.
[6] O. Haggag and S. E. Beltagy, "Plagiarism Candidate Retrieval Using Selective Query Formulation and Discriminative Query Scoring—Notebook for PAN at CLEF 2013", CLEF 2013 Evaluation Labs and Workshop – Working Notes Papers, Valencia, Spain, **(2013)**.
[7] A. K. Jayapal, "Similarity Overlap Metric and Greedy String Tiling at PAN 2012: Plagiarism Detection—Notebook for PAN at CLEF 2012. CLEF 2012 Evaluation Labs and Workshop – Working Notes Papers, Rome, Italy, **(2012)**.
[8] K. Williams, H. H. Chen, S. R. Chowdhury and C. L. Giles, "Unsupervised Ranking for Plagiarism Source Retrieval—Notebook for PAN at CLEF 2013", CLEF 2013 Evaluation Labs and Workshop – Working Notes Papers, Valencia, Spain, URL http://www.clef-initiative.eu/publication/working-notes, **(2013)**.
[9] Š. Suchomel, J. Kasprzak, and M. Brandejs, "Three Way Search Engine Queries with Multi-feature Document Comparison for Plagiarism Detection—Notebook for PAN at CLEF 2012", CLEF 2012 Evaluation Labs and Workshop – Working Notes Papers, Rome, Italy, **(2012)**.
[10] C. Grozea and M. Popescu, "Encoplot - Tuned for High Recall (also proposing a new plagiarism detection score)", CLEF 2012 Evaluation Labs and Workshop – Working Notes Papers, Rome, Italy, **(2012)**.
[11] D. A. R. Torrejón and J. M. M. Ramos, "Detailed Comparison Module In CoReMo 1.9 Plagiarism Detector—Notebook for PAN at CLEF 2012", CLEF 2012 Evaluation Labs and Workshop – Working Notes Papers, 17-20 September, Rome, Italy, **(2012)**.
[12] Y. Palkovskii and A. Belov, "Applying Specific Clusterization and Fingerprint Density Distribution with Genetic Algorithm Overall Tuning in External Plagiarism Detection—Notebook for PAN at CLEF 2012", CLEF 2012 Evaluation Labs and Workshop – Working Notes Papers, Rome, Italy, **(2012)**.
[13] T. Y. Liu, "Learning to Rank for Information Retrieval", Foundations and Trends in Information Retrieval, vol. 3, no. 3, **(2009)**, pp. 225-331.
[14] S. E. Robertson, "Overview of the okapi projects", Journal of Documentation, 53:3–7, 1997.S. E. Robertson. Overview of the okapi projects. Journal of Documentation, vol. 53, **(1997)**, pp. 3-7.
[15] K. Leilei, Q. Haoliang, W. Shuai, D. Cuixia, W. Suhong and H. Yong, "Approaches for Candidate Document Retrieval and Detailed Comparison of Plagiarism Detection", Notebook for PAN at CLEF **(2012)**.

Authors

Leilei Kong, born in 1979, ph. D. Candidate, associate professor. Her research interests include plagiarism detection, information retrieval, and natural language processing.

Copyright ⓒ 2016 SERSC

Copyright ⓒ 2016 SERSC