Detection of Paraphrastic Cases of Mono-lingual and Cross-lingual Plagiarism

Parth Gupta[1], **Khushboo Singhal**[2], **Prasenjit Majumder**[2], **Paolo Rosso**[1]

[1] Natural Language Engineering Lab - ELiRF
Department of Information Systems and Computation
Universidad Politécnica de Valencia, Spain
http://users.dsic.upv.es/grupos/nle
{pgupta,prosso}@dsic.upv.es

[2] IR-Lab, DA-IICT, India.
http://irlab.daiict.ac.in
{khushboo_singhal,p_majumder}@daiict.ac.in

Abstract

External plagiarism detection is a unique retrieval process where the algorithm has to provide an evidence of plagiarism if any for a suspicious section from the pool of source documents available. This paper focuses on paraphrasing involved in detection of plagiarism both from monolingual and cross-lingual aspect. In order to investigate the challenges in detection, we further analyse the performance of Vector Space Model(VSM) based external plagiarism detection system on PAN-PC-2011 corpus.

1 Introduction

In this study, we want to address the challenges involved in detection of paraphrastic cases of plagiarism both from a monolingual and cross-lingual perspectives. Cross-lingual plagiarism refers to the case of plagiarism where the suspicious and source documents are not in the same language. In order to do so, we investigate further the results obtained in these plagiarism cases by the VSM based external plagiarism detection system (Rao et al., 2011). This system participated in plagiarism detection track of PAN 2011[1] (Uncovering Plagiarism, Authorship, and Social Software Misuse) competition held in conjunction with the CLEF 2011[2] (Conference on Multilingual and Multimodal Information Access Evaluation), where it attained 6^{th} rank in external plagiarism detection category and 4^{th} rank in overall competition.

There are two types of automatic plagiarism detection: external and intrinsic. In the intrinsic case, the system needs to identify the plagiarized sections in the suspicious document without having any source documents to compare with. Usually these systems try to estimate the change in author writing profile. In case of external plagiarism detection, systems should identify the potential cases of plagiarism from the pool of source documents using Informational Retrieval techniques. On determination of candidate source documents, plagiarized sections should be identified using Natural Language Processing techniques. The plagiarized section can be a verbatim copy of the source sections or there might be paraphrasing involved. Paraphrase[3] is restatement of a text or passages, using other words. Paraphrase generation from the general viewpoint can be manual or automatic. There are three types of automatic paraphrase generation: corpora-based, statistical-based and dictionary or rule-based (Barreiro, 2011). The case of cross-lingual paraphrasing is due to automatic and/or manual translation for the comparison of the documents. (Barreiro, 2010) talks more about paraphrasing with automatic translation. Construction of such paraphrastic cases in the corpus is discussed in Section 3. We analyse the monolingual paraphrases of English and cross-lingual paraphrases for German and Spanish languages. This study aims to provide a knowledge-base for the future development and improvement of the external plagiarism detection systems by analyzing the paraphrases which are hard-to-detect. All the results are reported on PAN-PC-2011 dataset[4].

[1] http://pan.webis.de
[2] http://clef2011.org/
[3] en.wikipedia.org/wiki/Paraphrase
[4] Dataset:PAN-PC-2011, http://pan.webis.de

"Proceedings of ICON-2011: 9th International Conference on Natural Language Processing, Macmillan Publishers, India. Also accessible from http://ltrc.iiit.ac.in/proceedings/ICON-2011"

Section 2 describes the external plagiarism detection algorithm. In Section 3 we report the performance evaluation of the experiments carried for detecting plagiarism. We analyse the paraphrases in detail which are hard-to-detect in Section 4. In Section 5 we conclude the work.

2 External Plagiarism Detection

External plagiarism detection systems often deal with massive computational complexities. Ideal way to find the plagiarized chunks in the suspicious documents would be to compare each suspicious document to all source documents, but its computationally unreal. Hence the best way is to identify a set of documents which are very close to the suspicious document and can be a potential source of plagiarism. This step is referred to as a candidate retrieval (Potthast et al., 2009). There are various techniques reported for the same including, Fingerprinting (Zou et al., 2010), Winnowing, Vector Space Model (VSM) (Devi et al., 2010; Salton et al., 1975) based techniques. Using any of the mentioned techniques, a tractable number of candidate documents are retrieved and analyzed in detail. As described in (Potthast et al., 2010), the plagiarism detection can be seen as finding a 4-tuple $r =< r_{plg}, d_{plg}, r_{src}, d'_{src} >$ where r_{plg} is the detected plagiarized section in document d_{plg} which is plagiarized from from section r_{src} of source document d'_{src}. While the actual plagiarism case can be depicted as 4-tuple $s =< s_{plg}, d_{plg}, s_{src}, d_{src} >$ where s_{plg} is a plagiarized section in plagiarized document d_{plg} from the source section s_{src} of the source document d_{src}.

2.1 Algorithm

The algorithm can be broadly partitioned in four parts: Pre-processing, Candidate Retrieval, Detailed Analysis and Post-processing. Block diagram of the system is shown in Figure 1.

2.1.1 Pre-processing

It may be a case when two documents being compared are not in the same language. In such cases we have considered English as the language of comparison. We use two step strategy to translate all the documents which are not in English. First we identify the language of the document using Google Language Identifier[5] and then we translate all the non-english documents using Google Translator API[6] into english. Identification step saves computational overhead from processing all the documents in the translator.

2.1.2 Candidate Retrieval:

We use VSM based approach to select the candidate documents. Here we index all the source documents with stop-words removal and stemming. After that, we pass the suspicious document as a query to that index and a rank-list is generated. We consider top 250 documents as candidate documents or a set of documents for which Similarity Score is greater than 0.01, which ever occurs first. This strategy of involving two parameters for upper bound works good because there were many documents which were not at all plagiarized and this fact is revealed by the Similarity Score, for such documents the Similarity Score rapidly go below 0.01 on very early rank positions like top 10 or so and hence we save computational power by not analyzing all the 250 candidate documents for such suspicious document cases. Also we at least analyse top 20 documents for each suspicious document so that we may not miss any of those suspicious documents which have very small portion of plagiarism and so that Similarity Score is quite less. Here Similarity Score is referred by the following formula which is a typical Dot Product of two documents $d2$ the source document and q being the suspicious document.

$$cos\theta = \frac{d_2.q}{\|d_2\|.\|q\|} \quad (1)$$

2.1.3 Detailed Analysis:

There are lot of difficulties in detecting the plagiarism due to different levels of obfuscation involved. We believe that most of the time plagiarized chunk is not an exact copy at the same time the plagiarized chunk can not be completely different in words from the source. Our approach uses a window based similarity score to detect plagiarism. First of all we take a 7-word gram of the suspicious document and look for it in source document if it matches, we believe there can be a case of plagiarism because seven 'consecutive

[5]Google Language Identifier: http://code.google.com/p/language-detection/

[6]Google Language Translator: http://translate.google.com/

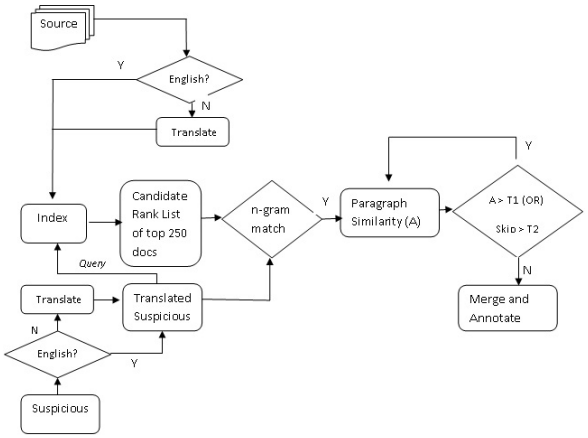

Figure 1: Block Diagram of External Plagiarism Detection System

words match' is a potential evidence. We use 7-word gram based on experience in (Gupta et al., 2010). Now from the matching point we take 25 words window in both suspicious and source document. The similarity score is calculated for them which is same as explained in last section. We do not consider stop-words in these windows because they unnecessarily increase the similarity score. We define a cut-off score of 0.50 to declare the passages having similarity score above it as plagiarized chunks. We keep on calculating the similarity scores for consecutive windows until 8 consecutive windows have similarity scores below cut-off. The reason behind keeping the 8 tolerance windows is that may help to handle false negatives due to obfuscation. False positives can be avoided by keeping the similarity score high. When one chunk is identified, the same process is repeated for the remainder of the suspicious document.

We merge the consecutive plagiarism cases if they are 500 characters apart. This helps in tackling obfuscation in one way and also helps to improve the granularity.

2.1.4 Post-processing:

Here we try to identify and remove the false positives due to VSM based approach. If a suspicious document annotated by our algorithm has no plagiarism cases of length greater than 160 characters, we consider that document plagiarism free.

3 Experiments

We tried the above mentioned algorithm on PAN-PC-2011 data to detect external plagiarism. It has 11093 suspicious and same number of source documents. Document distribution in the corpus is described in Table 1. We report the performance of the algorithm for particular categories of plagiarism which are hard-to-detect along-with *plagdet* score achieved. *Plagdet* is the overall score and is represented as below

$$plagdet = \frac{F}{\log_2(1 + granularity)} \quad (2)$$

Where F is standard F-Measure, i.e. harmonic mean of Precision and Recall. Precision refers to amount of actual plagiarism detected out of total detected plagiarism and recall refers to amount of actual plagiarism detected out of total actual plagiarism. These measures are calculated at character level. Granularity can be seen as, in how many parts one plagiarized section is detected and hence its ideal value is 1. *Plagdet* is computed by combining all three parameters as shown in the formula. More details about evaluation framework can be found in (Potthast et al., 2010).

Paraphrasing generation in the PAN-PC-2011 corpus can be mainly divided in two categories called *artificial* and *simulated*. Former cases are generated using heuristics like Random Text Operations, Semantic Word Replacements and POS Preserving Word-Shuffling while the latter cases

Document Statistics							
Document Purpose		Plagiarism per Document			Document Length		
source documents	50%	hardly	(5%-20%)	57%	short	(1-10 pp.)	50%
suspicious documents		medium	(20%-50%)	15%	medium	(10-100 pp.)	35%
- with plagiarism	25%	much	(50%-80%)	18%	long	(100-1000 pp.)	15%
- without plagiarism	25%	entirely	(>80%)	10%			

Table 1: Document statistics in the PAN-PC-2011 (Potthast et al., 2011)

Plagiarism Cases				
Obfuscation		Case Length		
none	18%	short	(<150 words)	35%
paraphrasing		medium	(150-1150 words)	38%
- automatic (low)	32%	long	(>1150 words)	27%
- automatic (high)	31%			
- manual	8%			
translation (de, es to en)				
- automatic	10%			
- automatic + manual correction	1%			

Table 2: Statistics and distribution of plagiarism cases in PAN-PC-2011 (Potthast et al., 2011)

are generated using Amazon Mechanical Turk (Barr and Cabrera, 2006). There are different levels of *artificial* plagiarism cases with number of operations involved and range of affected phrases, which can be classified into three categories: low, medium and high obfuscation. Corpus also involves cross-lingual plagiarism cases with *artificial* and *simulated* paraphrasing. More details of the cases generation can be found in (Potthast et al., 2010). Distribution of obfuscation in plagiarism cases is listed in Table 2 (Potthast et al., 2011). Though the corpus contains plagiarism cases with obfuscation levels none, low, medium and high as well as simulated cases and translated cases, we focus the analysis on simulated and artificial cases with high obfuscation for both monolingual and cross-lingual cases because of their paraphrastic richness. Results are shown in Table 3.

4 Analysis

Poor performance of the system on hard-to-detect plagiarism cases encourages us to closely look at the individual cases in order to make future system robust. Hence some of the sections involved in plagiarism cases are extracted to understand the paraphrasing better. In the subsections below, we present case-wise analysis for monolingual high obfuscated artificial and monolingual simulated cases, which can be seen as automatic vs. manual paraphrasing along-with cross-lingual cases with automatic and manual translation. Reasons behind successful detection of plagiarized phrases as well as undetected phrases are also discussed.

4.1 Monolingual High Obfuscated Paraphrasing

Table 4 shows the case of high obfuscated artificial paraphrasing which involves high amount of text operations described in the previous section. Such cases can be grammatically incorrect which can be verified from the case shown in the Table 4.

We observe that, VSM based approach provides an ideal platform to detect the plagiarism cases of such type because of its leverage to match the sections as bag of words model. We also noticed, if the artificial case has less than 50% of words in the 25 words window changed, our system detects the case because of similarity threshold being set to 0.50. Undetected part of the reported case gives an evidence of high obfuscation with words replaced beyond 50% of limit and hence not detected. It is also observed, this parameter setting of 25 words window and 0.50 similarity score fetches many false positives which gives a scope for further parameter tuning. To handle semantic word replacements one can incorporate dictionary or thesauri of the language if available, which might help to raise threshold in order to minimize false positives.

4.2 Monolingual Simulated Paraphrasing

In contrast to the artificial cases, simulated cases are more realistic because they uphold the grammatical construction of the sentences. In this type of cases, sentences are rephrased to express the

Paraphrase Type	Plagdet Score	Recall	Precision	Granularity
Monolingual High Obfuscated	0.0314059	0.0188260	0.1819626	1.1235452
Monolingual Simulated	0.0524298	0.0293390	0.3780321	1.0541872
Cross-lingual High Obfuscated	0.0846094	0.0864208	0.1434008	1.4193989
Cross-lingual Simulated	0.0320123	0.0281610	0.0389451	1.0294118
Overall	0.1990889	0.1618067	0.4541152	1.2949292

Table 3: Results for External Plagiarism Detection on PAN-PC-11 corpus for hard-to-detect plagiarism categories as well as for overall corpus

Source Section	Plagiarized Section
.. **Every barrister in England must be a member of one of the four ancient societies called Inns of Court, viz. Lincoln's Inn, the Inner and Middle Temples, and Gray's Inn, and in Ireland, of the King's Inns.** The existence of the English societies as schools can be traced back to the 13th century, and their rise is attributed to the clause in Magna Carta, by which the Common Pleas were fixed at Westminster instead of following the king's court..	.. **The lawyer of England must be the areopagite in one for a four ancient civilization wound Hostel before Lawcourt, viz. Lincoln'element Caravansary, the outer and late Feature, and Grayness'element Hotel, and of Eire, of a Rex'element Caravansary.** The actuality among the english society in academy may be split away in any 13th period, and their fall is delegate of the deductible of Magna Carta, in which the individual Supplication were fixed in Westminster besides to keep the queen'element tribunal..

Table 4: Case of monolingual high obfuscated artificial plagiarism from PAN-PC-2011. Detected part by the algorithm is shown in bold letters.

Source Section	Plagiarized Section
... **System, punctuality, industry, belong to the Dead-letter Office. It seems to embrace every other branch of business, and, as I have shown you, even to know how to treat such unwelcome guests as a nest of live serpents.HOW MOTHER ROBIN CALLED A NEW MATE. BY E. JAY EDWARDS.** A friend of mine has a robin's nest that he guards with very great care, and about which he tells a story to all the young and old people who call upon him. "There is a romance" he says, as he shows you the nest, "about this, and if you want to hear it, I will tell it to you."...	... **In the Dead-Letter Office there is system, industry and punctuality. Apparently it embraces every other branch of business, and as you have seen, treating an unwelcome guest such as a nest of live serpents is in a must-know basis.HOW MOTHER ROBIN CALLED A NEW FRIEND. BY E.JAY EDWARDS** "Something special is held," he says dreamingly as he shows you the nest, "by this, if you would like to know what it is, I will tell you." ...

Table 5: Case of simulated plagiarism from the PAN-PC-2011 dataset generated using Amazon's Mechanical Turk. Detected part by the algorithm is shown in bold letters.

same message rather than replacing the words mechanically.

We observe, sometimes in order to rephrase the section Amazon's Mechanical Turk adds more sentences in-between or shuffles the sentences. In such cases, due to text insertion, the algorithm looses an overlap for the sections comparison which it gained from a 7-word gram match. This fact can easily be verified from the case shown in Table 5 where the paragraphs *"There is a romance..."* in the source section and *"Something special is held..."* in the suspicious section could have been detected if the paragraph *"A friend of mine..."* in the source section had not been inserted. We believe, increasing the window size may help in such cases.

4.3 Cross-lingual High Obfuscated Paraphrasing

Table 6 shows the case of artificial plagiarism generated using cross-lingual high obfuscation. In our detection system we use Google Translator API which seems to be giving the very close translation to the generated one but due to some character

Original Source Section	Cross-lingual High Obfuscated Suspicious Case	Translated Source Section by Algorithm
... THIERS.–Pues no la haba de ver! En el vagn reservado estaba, bien abrigadita, sin decir esta boca es ma, y tan contenta que echaba lumbre por los ojos... ROSALA.–¿Y tuviste paciencia para presenciar tal escándalo?... ¡Con que no la puede hacer su mujer porque es una... y la hace su querida...! Estoy volada... Ignominia tan grande en nuestra familia, en esta familia honrada y ejemplar como pocas, me saca de quicio... =(Mirándole con fuerza.)= ¿Y tú no dijiste nada?, ¿aguantaste que en tus barbas?...	... **THIERS .– As he had not see! In the car was quiet, well sheltered at, not saying This mouth is mine, and so happy that threw light on the eyes.. ROSALIA .– And you had the patience to witness such a scandal? ... With that you can not do his wife because she is a ... and make your darling ...! I am blown ... So great shame** our family, this family honored and exemplary few, drives me crazy ... = (Looking at you hard.) = "And do not say anything?," put up with in you...	... **THIERS .– As he had not see! In the car was quiet, well sheltered at, not saying a peep, and so happy that threw light on the eyes ... ROSALIA .– And you had the patience to witness such a scandal? ... With that you can not make your woman because she is a ... and make your darling ...! I am blown ... So great shame** ...

Table 6: Case of cross-lingual high obfuscated artificial plagiarism from the PAN-PC-2011 dataset. Detected part by the algorithm is shown in bold letters.

level differences, API acted differently for some sentences making cases even harder to detect. If the translator used for detection is different from the translator used for generation, there might be change in sentence formation. We observe that in cross-lingual cases there are some translation conflicts like *"mujer"* word of spanish language is translated to *"wife"* in generated case while translated to *"woman"* in our algorithm using Google Translate. Linguistically both are correct translation with respect to the context and VSM based approach detects it.

4.4 Cross-lingual Simulated Paraphrasing

Table 7 shows the case of cross-lingual simulated plagiarism. In such cases, when we translate the non-english document using Google Translate, paraphrasing occurs. This fact can be observed in the example shown in the Table 7. The translation is less near to the generated case in comparison to artificial high obfuscated cases. Such types of plagiarism cases are detected if the discussed limit of 50% words are obeyed. There is paraphrasing which is detectable like *"will see me"* and *"will find me"*, *"nuts"* and *"walnut trees"*, *"hundred steps"* and *"hundred yards"* etc. Such paraphrasing arises due to automatic paraphrasing generation on one hand and translating the document using other Machine Translation services like Google Translate on the other hand.

4.5 Other Challenges

The performance of the system is determined based on the annotated offset and length. We noticed that most of the translated source documents using Google Translate do not preserve offset information as in the original source documents. Hence there is a big difference between the offset of a sentence in the original source document and offset of the same sentence in the translated document. This deteriorates the performance of the system in terms of both precision and recall.

Another important issue from performance of algorithm perspective is the quality of candidate retrieval phase. Recall of the system is always upper-bounded by the rightly retrieved candidate documents. We observe that our candidate retrieval phase still lags to pull all the source documents used to plagiarise the suspicious document particularly when plagiarism length is very small. In case of simulated and high obfuscated artificial plagiarism cases many of the words are replaced or rephrased and hence the VSM score for such source documents go below threshold and hence they are not retrieved too.

We believe, automatic machine translation of all the documents in other languages to language of comparison is not the ideal way to handle cross-lingual plagiarism detection for many reasons. If we consider World Wide Web (WWW) as the source, it is computationally unreal to translate all the documents. Moreover, automatic machine translation always incurs some semantic informa-

Original Source Section	Cross-lingual Simulated Suspicious Case	Translated Source Section by Algorithm
...Morgen ist Dein Geburtstag; ich muss Dich sehen, zum letzten Male sehen. Heute abend vor dem Tore findest Du mich in dem kleinen Waeldchen, unter den Nussbaeumen, etwa hundert Schritte vom Wege, bei der kleinen Kapelle rechts. Wenn Du mir einiges Geld zu meiner Huelfe mitbringen kannst, so wird Dir es Gott vergelten...	... **Your birthday is tomorrow, and must see you, see the last time. In the evening, near the gate you'll see me in a small forest, including nuts, a hundred steps the road, the small chapel on the right. If you give me some money can bring to my aid, so that God will repay**...	... **Your birthday is tomorrow, I must see you, for the last time . see Tonight, before the gates you can find me in the little grove, under the walnut trees, about a hundred yards from the road, turn right at the small chapel. If you give me some money to can bring my help, then you will repay God**...

Table 7: Case of cross-lingual simulated plagiarism from the PAN-PC-2011 dataset. Detected part by the algorithm is shown in bold letters.

tion loss which in turn affects the detection. The cases shown in the Sections 4.3 and 4.4 have very close translation to the actual plagiarism cases because the same automatic machine translator is used for the case generation. The performance will heavily be affected if the automatic translation service used for detection is different from the service used for plagiarism.

The automatic plagiarism detection systems merely identify the text which is very close to the source text in terms of words or semantics or both. These systems heavily rely on basic and advanced text comparison techniques with a trade off with computational complexity. Post-processing phase of these systems involve some knowledge based operations like removal of cases which involve proper citation, yet they are in their preliminary states. These steps usually help in identifying false-positives. At the end, they are meant to assist human by providing closest evidences and based on them a human has to decide whether it is plagiarism or not.

5 Conclusions and Future Work

This study helps in understanding the nature of automatic paraphrasing involved in detecting plagiarism for monolingual as well as cross-lingual artificial and simulated cases in text. VSM based system is a good choice for the detailed analysis phase because of its capability to match the sections irrespective of the orderings among words. This approach can be more effective if clubbed with any synonym addition strategy using thesauri, dictionary or wordnet. In future we want to tune parameters of the system to incorporate the knowledge gained from this study. We also want to see the performance change after including dictionary or thesauri based word addition.

6 Acknowledgment

The work of the researchers of the Technical University of Valencia has been done in the framework of the VLC/CAMPUS Microcluster on Multimodal Interaction in Intelligent Systems and it has been partially funded by the European Commission as part of the WIQ-EI IRSES project (grant no. 269180) within the FP 7 Marie Curie People Framework, and by MICINN as part of the Text-Enterprise 2.0 project (TIN2009-13391-C04-03) within the Plan I+D+i.

References

Jeff Barr and Luis-Felipe Cabrera. 2006. Ai gets a brain. *ACM Queue*, 4(4):24–29.

Anabela Barreiro. 2010. *Make It Simple with Paraphrases*. LAP Lambert Academic Publishing.

Anabela Barreiro. 2011. Spider: A system for paraphrasing in document editing and revision - applicability in machine translation pre-editing. In *CICLing, Springer-Verlag, LNCS(6609)*, pages 365–376.

Sobha Lalitha Devi, Pattabhi R. K. Rao, R. Vijay Sundar Ram, and A. Akilandeswari. 2010. External plagiarism detection - lab report for pan at clef 2010. In *Notebook Papers of CLEF 10 Labs and Workshops*.

Parth Gupta, Sameer Rao, and Prasenjit Majumder. 2010. External plagiarism detection: N-gram approach using named entity recognizer - lab report for pan at clef 2010. In *Notebook Papers of CLEF 10 Labs and Workshops*.

Martin Potthast, Benno Stein, Andreas Eiselt, Alberto Barrón-Cedeño, and Paolo Rosso. 2009. Overview

of the 1st International Competition on Plagiarism Detection. In Benno Stein, Paolo Rosso, Efstathios Stamatatos, Moshe Koppel, and Eneko Agirre, editors, *SEPLN 2009 Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN 09)*, pages 1–9. CEUR-WS.org, vol. 502.

Martin Potthast, Benno Stein, Alberto Barrón-Cedeño, and Paolo Rosso. 2010. An evaluation framework for plagiarism detection. In *COLING*, pages 997–1005.

Martin Potthast, Andreas Eiselt, Alberto Barrón-Cedeño, Benno Stein, and Paolo Rosso. 2011. Overview of the 3rd international competition on plagiarism detection. In *Notebook Papers of CLEF 11 Labs and Workshops*.

Sameer Rao, Parth Gupta, Khushboo Singhal, and Prasenjit Majumder. 2011. External & intrinsic plagiarism detection: Vsm & discourse markers based approach - notebook for pan at clef 2011. In *Notebook Papers of CLEF 11 Labs and Workshops*.

Gerard Salton, A. Wong, and C. S. Yang. 1975. A vector space model for automatic indexing. *Commun. ACM*, 18(11):613–620.

Du Zou, Wei jiang Long, and Zhang Ling. 2010. A cluster-based plagiarism detection method - lab report for pan at clef 2010. In *Notebook Papers of CLEF 10 Labs and Workshops*.