BAR CHART PLAGIARISM DETECTION

MOHAMMED MUMTAZ MOHAMMED SALIH

UNIVERSITI TEKNOLOGI MALAYSIA

BAR CHART PLAGIARISM DETECTION

MOHAMMED MUMTAZ MOHAMMED SALIH

A dissertation submitted in partial fulfillment of the
requirements for the award of the degree of
Master of Science (Computer Science)

Faculty of Computer Science and Information Systems
Universiti Teknologi Malaysia

JANUARY 2013

This dissertation is dedicated to my family for their endless support and encouragement.

ACKNOWLEDGEMENT

"In the Name of Allah, Most Gracious, Most Merciful"

First and foremost, *Alhamdulillah*, it is with the assistance and grace of *Allah Almighty* that I was able to finish this dissertation.

I would like to express my sincere appreciation to my supervisor Prof. Dr. Naomie Salim for her great advice and generous help during the period of my study and also, who had the patience and wisdom to guide me in order to overcome all the academic obstacles that I faced during my study. My overwhelming gratitude to my evaluators, I am also grateful for their helpful suggestions.

A special thanks to my parents and my siblings for their unlimited moral support, to everyone in my extended family, and for their lessons on how to be patient and strong. I thank them very much for always being there for me and I ask Allah the almighty to grant them Paradise.

Last but not least, I would like sincerely to thank all the lectures, staff, friends and my fellow postgraduate students for their emotional support and cognitive, thanks for all the care and concern. I wish you more and brighter success in this world and the Hereafter.

ABSTRACT

Plagiarism can be considered one of the electronic crimes and intellectual thefts, which has become one of educational challenges of research institutions. One form to represent quantitative information is charts such as line and bar chart, which can formulate the information in info-graphic form. The extraction of features of bar chart is an essential process to get the data from images. Some techniques presented by researchers focused on the graphical part rather than text itself, such as Hough Transform and Learning Based method. In this study, ten features of bar chart images are utilized to detect and find the proportion of similarity between the charts. Some of these features can be directly extracted by OCR, while others demand finding the relationship between the text part and the graphic part to extract the data such as the real values for each bar in images. The new technique which introduced in this research can extract three values of each bar namely Start, End and Exact values depending on horizontal and vertical lines of the bar chart image. In addition, the Word 2-gram and Euclidean distance methods are used to detect and find the plagiarism. Experimental results show the ability of the system to detect plagiarism for ten possible patterns of bar chart plagiarisms. The performance of the system is evaluated depending on overlapping features and precision and recall. The experimental results show the ability of the system to detect not only copy and paste data of bars, but also restructuring and summarization of captions of image as well as modifications to data of bar chart images, such as swapping among bars, changing colors and changing scales of bar chart images.

ABSTRAK

Penciplakan merupakan salah satu jenayah elektronik dan kecurian intelek merupakan salah satu cabaran dalam bidang pendidikan dan penyelidikan. Penggunaan carta seperti carta bar dan carta garis dalam penulisan merupakan satu bentuk penyampaian maklumat secara info-grafik. Pengekstrakan ciri-ciri penting dari carta bar adalah satu proses untuk menghasilkan perwakilan imej. Teknik-teknik yang telah digunapakai oleh penyelidik-penyelidik lain adalan seperti *Transformasi Hough* dan *Berasaskan Pembelajaran* lebih mem fokus kepada bahagian grafik daripada bahagian teks. Dalam kajian ini, sepuluh ciri-ciri perwakilan imej carta bar digunakan untuk mengesan kadar persamaan antara imej. Ada di antara ciri-ciri tesebut yang boleh diekstrak menggunakan *OCR*, manakala selebihnya melalui hubungan di antara bahagian teks dan bahagian grafik seperti nilai setiap bar dalam imej. Teknik baru yang dicadangkan di dalam kajian ini boleh mengekstrak tiga nilai untuk setiap bar iaitu nilai mula, nilai akhir dan nilai yang tepat bergantung kepada garis menegak dan mendatar bagi imej carta bar tersebut. Seterusnya, kaedah *Word 2-gram* dan *jarak Euclid* digunakan untuk mengesan kadar penciplakan. Hasil ajikaji menunjukkan bahawa sistem ini berkebolehan mengesan penciplakan bagi sepuluh bentuk penciplakan carta bar. Prestasi sistem diukur dan dinilai dengan melihat kepada ciri-ciri yang bertindan dan juga ukuran *kepersisan dan perolehan kembali*. Kesimpulannya, hasil eksperimen menunjukkan bahawa sistem yang dicadangkan, berkebolehan mengesan penciplakan data pada carta bar, termasuk penstrukturan semula dan pada imej dan pengubahsuaian lain seperti penukaran antara bar, penukaran warna dan skala pada imej carta bar.

TABLE OF CONTENT

LIST OF TABLES

TABLE NO.	TITLE	PAGE

LIST OF FIGURES

FIGURE NO. **TITLE** **PAGE**

FIGURE NO.	TITLE	PAGE

CHAPTER 1

INTRODUCTION

1.1 Overview

In recent years, the significant development in digital revolution represented by digital libraries and World Wide Web can be considered as one of the main reasons for dramatical growth in the appearance of plagiarism. As, most resources are available in digital format, it has become easy for the plagiarist to utilize or take other people's works without referencing or quoting the owner for his/her work. So, plagiarism can be considered as one of the electronic crimes and intellectual thefts of efforts of people (Ali *et al.*, 2011).

Many students and researchers may plagiarize other people's works to obfuscate that this work represents their own works. Therefore, plagiarism detection in academic area has become one of the most important educational challenges that the research institutions, universities and even schools are facing. There are a lot of studies to detect plagiarism of programming code and text and attempt to cover various types of plagiarism.

Plagiarism detection has begun since 1970s, for detecting the rate of plagiarism of programming code which is written by some programming language such as C and Pascal, (Alzahrani *et al.*, 2012).

The plagiarism code program or algorithm can be defined as the program that is created or reproduced from another program to identify the plagiarist to the owner of this work. Many researchers presented studies to detect the plagiarism of computer algorithm and program code. (Ottenstein, 1976), suggested the algorithmic approach to detect plagiarism for homework that is presented by students. Moreover, later studies introduced several levels of plagiarism that can be used in algorithms and program code, these levels explain the types and patterns of plagiarism, (Parker and Hamblen, 1989).

While researchers tend to detect the plagiarism in natural language since 1990s, the computerized or statistical approaches are utilized to detect plagiarism in natural language. These studies paved the way for studying the mechanism of copy detection in digital resources documents, (Alzahrani, *et al.*, 2012). The techniques that are used in natural language are based on several factors, which are grammar-based, semantic-based and grammar semantics hybrid method, (Ali, *et al.*, 2011).

The grammar-based method is one of restricting techniques to detect the plagiarism. This type of method analyzes the sentences based on grammatical structure. Therefore, grammar-based method can be efficiently used to detect the exact copy of text. In contrast, the modified text which uses synonyms and rewriting of some of sentences requires other techniques to effectively detect the plagiarism. While semantic-based method utilizes vector space model to calculate the similarity of text. Semantic-based technique also has drawbacks when used to detect plagiarism of partial text in documents, but when used for the whole text in document, it is efficient. In spite of these disadvantages of grammar-based and semantic-based method, these methods can be considered as one important approach to detect plagiarism.

However, the method that overcomes on all disadvantages of previous techniques is grammar semantics hybrid method. Therefore, it can be considered as one of the most important techniques to detect the text plagiarism, (Ali, *et al.*, 2011; Bao Jun-Peng, 2003).

(Alzahrani, *et al.*, 2012), introduced new taxonomy which explains the concepts for various types and patterns of text plagiarism. They divided the plagiarism into two main parts which are, literal plagiarism and intelligent plagiarism. Each part consists of several subparts which cover all possibilities of text plagiarism. Additionally, the techniques and methods used to detect text plagiarism are introduced.

On the other hand, the representation of quantitative information can formulate in info-graphic form by using figures, charts and even tables. The information that are exhibited in charts, figures and tables include results of experiments, framework and statistical information. The data and information that are represented in charts and figures are in homogenous form. The information in charts and figures can be formulated by using various shapes such as; pie chart, bar chart, 2-D and 3-D plot figure, (Huang *et al.*, 2007).

However, tables are one of the important methods to present relational information in an efficient and compact manner whereas; the documents nowadays include different kinds of tables, (Wang *et al.*, 2002). The plagiarism of charts, figures and tables can be defined by taking info-graphic of data from another work without quotation and citation to manifest the plagiarist as the ownership of this work. The visual information and data in chart and figure are represented as image. Therefore, the major challenge is how to extract the data and information from charts and figures in order to detect the proportion of plagiarism.

There are several types of charts such as; pie charts, bar charts and line charts. In this study, the bar chart will investigate how to detect the rate of plagiarism for data and information that are available inside the bar chart image.

1.2 Problem Background

This section provides an overview of related works which were introduced by researchers. There are no specific studies that introduced plagiarism detection of bar chart images. However, there are some studies which can extract the features and data from chart images.

There are several studies which introduced by researchers presented classification methods of chart images. The classification method which based on multiple-instance learning is introduced by (Huang, *et al.*, 2007), while (Savva *et al.*, 2011) introduced ReVision system which includes three concatenated major stages namely; classification, extraction and redesigned chart images. The input of their ReVision system is bitmap chart images. Then, the ReVision system automatically generates redesigned visualization chart image as output. In classification stage, there are two types of features which can be considered as main factors for ReVision system to classify the chart images, namely; low-level image features and text-level features. Meanwhile, in extraction stage, they focused their study on two types of charts which are, pie and bar charts. They presented some techniques and methods that can be used for extracting the data and graphical marks from chart images. In this stage, there are three levels of process in order to extract features from chart images. Firstly, the slices of pie and bar chart are situated in order to encode the data. Secondly, applied heuristics method which joins between marks and related elements like for example, axes with labels. Finally, utilize data and information obtained from previous steps in order to extract the table of data values. This information of chart images will be utilized in the redesigned stage in order to automatically generate redesigned visualization.

The understanding and recognition of chart image require the preprocessing and extracting the data and information. There are two types of methods which deal with chart image that were introduced by researchers, which either deal with electronic chart directly or deal with chart images after converting them into raster images. The studies that were introduced to deal directly with electronic chart are by (Carberry *et al.*, 2003; Elzer *et al.*, 2005); while other studies deal with chart images by converting them into raster images, this type of studies was introduced by, (Huang *et al.*, 2004; Yokokura and Watanabe, 1998; Zhou and Tan, 2000; Zhou and Tan, 2001).

On other hand, there are several techniques to extract the features from chart images. One of these techniques applied on two-dimension plot chart images, while other techniques applied on bar chart images. Some studies which introduced by researchers used Hough Transform technique as approach to extract the features of bar chart images (Zhou and Tan, 2000), while other studies based on the edges of bars to extract the features (Huang, *et al.*, 2004; Huang and Tan, 2007). The learning-based method introduced by (Zhou and Tan, 2001) to recognize the chart images. The features of bar chart images can extract by description of the bars such as height and width for each bar, this technique applied on statistical images to calculate the similarity (Hassan and Khatib, 2007). Whereas, other technique focused on geometric features rather than data and information of scientific bar chart images (Yang *et al.*, 2006). Table 1.1 shows the techniques of bar chart features extraction which introduced by researchers

Table 1.1: The techniques for extraction the features of bar chart images

No.	Techniques	Description	References
1.	Hough Transform Techniques	This technique can detect the boundaries tracing of bars.	(Zhou and Tan, 2000),
2.	Learning Based method	It focused on graphical parts rather than textual parts in bar chart image.	(Zhou and Tan, 2001)
3.	Model-based approach	It can detect and recognize the chart image based on edeges three types of chart images namely, line, pie and bar chart.	(Huang, *et al.*, 2004; Huang and Tan, 2007)
4.	Features description method	Describe the features of bar chart such as width and height to find similarity of statistical images	(Hassan and Khatib, 2007).

The two-dimension plot is one of the types of chart which is more popularly used by researchers and authors to exhibit experimental results in financial reports and scientific articles. This type of chart was introduced by (Kataria *et al.*, 2008). They presented an automatic method and technique to extract data and text from 2-D plot images; then, storing this information in databases as metadata. The interpretation of the features in the chart can easily be illustrated by humans, in contrast to the interpretation of the chart by computer which requires more analysis to obtain the data and information from chart images. Their method divided the 2-D plot into three regions. Firstly, the region of X-axis which includes label of X-axis and numerical unites. Secondly, the region of Y-axis which also consists of label of Y-axis and numerical units. Finally, curve region which includes legend text and data points in curve-fitted plots.

The Optical Character Recognition (OCR) techniques can use to extract the text features of chart images. The OCR tool was used by (Kataria, *et al.*, 2008) to detect label text block for X-axis and Y-axis after removing noise, which is one of factors that influence detecting text blocks. The features of 2-D plot will be extracted after having segmented the chart into three regions. The features that have been extracted by their algorithm include three parts; firstly, the X-axis and Y-axis of 2-D plot chart, secondly, the ticks that are located on the two axes and finally, text label that is associated with ticks. The OCR tool also was used by (Yang, *et al.*, 2006) to extract the text features in order to calculate the similarity of chart image.

Hough Transform technique is one of the techniques which used to detect and extract the features of bar chart image. This technique was introduced by (Zhou and Tan, 2000). The Hough Transform can detect the boundaries tracing of bars in bar chart images. Their system consists of three main stages, which are preprocessing, detection and recognition stage. The bar chart image segments to separate a text image and a graphic image in preprocessing stage. While in detection stage, they newly developed for Modified Probabilistic Hough Transform (MPHT) algorithm to detect the parallel lines of bar chart images. Finally, in recognition stage they reconstructed the patterns of bars and text primitive grouping. Their system can read and recognize different types of bar chart images, such as hand-drawn and skew bar chart images.

One of techniques for char recognition and understanding is Learning-Based approach which presented by (Zhou and Tan, 2001). The Learning-Based method focused on graphical parts rather than textual parts in bar chart images. They divided the Learning-Based approach into two main kinds, which are Learning-Based approach based on neural network, and Learning-Based approach based on Hidden Markov Model (HMM). The HMM can define a probabilistic model for time serious data. The HMM in their system generated various categories of chart from training images data. While in neural network part, they modified Back-Propagation algorithm in order to increase the speed of convergence. They extracted the features of bar chart images by novel principle approach. The features of bar chart images in their system divided into two types, which are principle invariant features and

principle variant features. The principle invariant features extracted for training while principle variant features extracted for information extraction

The Model-Based approach is one of the extraction and recognition techniques for chart images which introduced by (Huang, *et al.*, 2004; Huang and Tan, 2007). The Model-Based approach focused on edges to detect and extract the features of bar chart images as well as pie and line chart images. Whereas, there is other approaches that introduced by some researchers focused on extracting geometry features rather than data of chart image such as, (Yang, *et al.*, 2006). They introduced a semi-automatic system of scientific chart images that demands interaction by the user to generate vector of chart in order to use later for building chart image recognition system.

The features description of bar chart images is one of extracting techniques for data and information from chart images. This Techniques was introduced by (Hassan and Khatib, 2007). They presented a prototype which provides searching for similarity of statistical images. The features extracting were a major challenge to obtain the information and data from statistical images. Then, these features which are presented in databases are as metadata in order to be able to compare with the query. In statistical images, there are two kinds of features which are; global features such as, color histogram and local features such as, shape object. Their proposed system analyzes common features in the charts in order to calculate the similarity of statistical images. The important features for bar chart image which defined and used in their system are the number of bars, height and width for each bar, sequence of color and textual annotation.

In this study, the bar chart images will be used for creating the bar chart plagiarism detection system. The bar chart plagiarism system should have the ability of extracting the data and text information from bar chart images. Then, storing this information and data in database as metadata in order to compare it later with query image for detecting the proportion of plagiarism.

1.3 Problem Statement

The plagiarism of bar chart can be formulated in several forms as well as the plagiarism can occur completely or partly of the bar chart image. The formulating plagiarism of bar chart image can be either, copying the image without any modification, or changing and modifying some information of bar chart image such as, swapping among bars to deceive about the ownership of the work.
Therefore, the hypothesis of this study can be stated as:

"What techniques can be used to the plagiarism of bar chart images"

1.4 Dissertation Aim

The aim of the study is to develop techniques that can be used to detect the plagiarism of bar chart images.

1.5 Dissertation Objectives

The main objectives of this study are:

1. To identify and describe different techniques and methods of extraction the features and data from bar chart images.
2. To design techniques to detect bar chart plagiarism.
3. To validate the proposed plagiarism detection system based on precision and recall.

1.6 Scope of the Study

The scopes of this study are explained below:

1- Develop the program of proposed system Using MATLAB r2010a.
2- The type of graphic plagiarism that investigates in this study is Bar chart image.
3- Extract the text from the bar chart images by using OCR Tools manually, which can extract the text that is written on a horizontal line.
4- The dataset of bar chart that is prepared in this proposed system is commonly used for 2-D and some types of 3-D images.
5- Using synonyms for features of bar chart images are not investigated in this study.
6- Translation from any other languages to English has not investigated in this study.
7- In this study, generate other bar chart image for same data are investigated such as convert horizontal bar image into vertical one and vice versa, while generate other shapes for features of bar chart image have not investigated.
8- Using word 2-gram and Euclidean distance methods to detect and find the plagiarism.

1.7 Significance of the Study

The bar chart images are used to present the information and data in info-graphic form. The bar chart image can exhibit data of experiment results, statistical information or comparison among results. This type of quantitative information can quote from other work without citation. This study will introduce a bar chart plagiarism detection system which has the ability to detect the rate of plagiarism. The investigation of bar chart image to find an efficient technique for extracting the information and data from images will be presented. Besides, representing this data in database as metadata in order to compare later with query. Finally, the performance of the proposed system will be investigated by using evaluation approaches.

1.8 Dissertation Organization

This dissertation includes five chapters. The introduction, problem background, problem statement, aim of the study, objectives, scope and contribution of this research are presented in chapter One. The taxonomy of chart plagiarism, taxonomy of figure plagiarism, and taxonomy of table plagiarism are presented. Also, the patterns of plagiarism for bar chart that covered, the bar chart plagiarism detection techniques, and the Optical Character Recognition (OCR) are explained in chapter Two. In chapter Three, the methodology is discussed, it consists of nine phases started from planning research and introducing chart plagiarism framework until reached to investigation and evaluation the system. Chapter Four shows the experimental results of ten possible patterns of bar chart plagiarism, as well as the evaluation for each pattern and overall performance of system. Finally, Chapter Five indicates the conclusion, finding and suggestion for future work.

REFERENCES

Ali, A. M. E. T., Abdulla, H. M. D. and Snasel, V. (2011). Survey of Plagiarism Detection Methods. Paper presented at the *Proceedings of the 2011 Fifth Asia Modelling Symposium*.

Alzahrani, S. M., Salim, N. and Abraham, A. (2012). Understanding Plagiarism Linguistic Patterns, Textual Features, and Detection Methods. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 42(2), 133-149. doi: 10.1109/tsmcc.2011.2134847

Bao Jun-Peng, S. J.-Y., Liu Xiao-Dong, Song Qin-Bao. (2003). A Survey on Natural Language Text Copy Detection. *Journal of Software*, 14(10), 1753-1760.

Basile, C., Matematica, D., Benedetto, D., Caglioti, E., Cristadoro, G. and Esposti, M. D. (2009). Caglioti E.: A plagiarism detection procedure in three steps: selection, matches and 'squares. Proceedings of the 2009 *SEPLN 2009 Workshop on Uncovering Plagiarism, Authorship and Social Software Misuse (PAN 09*. 2009. 1-9.

Carberry, S., Elzer, S., Green, N., McCoy, K. and Chester, D. (2003). Understanding Information Graphics: A Discourse-Level Problem. Proceedings of the 2003 *IN PROC. OF SIGDIAL 2003. ASSOC. FOR COMPUTATIONAL LINGUISTICS*. 2003. 1-12.

Chaudhuri, B. B. and Pal, U. (1998). A complete printed Bangla OCR system. *Pattern Recognition*, 31(5), 531-549. doi: 10.1016/s0031-3203(97)00078-2.

Corezola Pereira, R., Moreira, V. and Galante, R. (2010). A New Approach for Cross-Language Plagiarism Analysis. In M. Agosti, N. Ferro, C. Peters, M. Rijke & A. Smeaton (Eds.), *Multilingual and Multimodal Information Access Evaluation* (Vol. 6360, pp. 15-26): Springer Berlin Heidelberg.

Doraisamy, S. (2004). *Polyphonic Music Retrieval: The N-gram Approach.* PHD, University of London, London.

Doraisamy, S. (2004). *Polyphonic Music Retrieval: The N-gram Approach.* PHD, University of London, London

Eikvil, L. (1993). Optical Character Recognition (OCR).

Elhadi, M. and Al-Tobi, A. (2009). Duplicate Detection in Documents and WebPages using Improved Longest Common Subsequence and Documents Syntactical Structures. Proceedings of the 2009 *International Conference on Computer Sciences and Convergence Information Technology*. 2009. Seoul, Korea, 679-684.

Elzer, S., Carberry, S., Zukerman, I., Chester, D., Green, N. and Demir, S. (2005). A probabilistic framework for recognizing intention in information graphics. Paper presented at the *Proceedings of the 19th international joint conference on Artificial intelligence*, Edinburgh, Scotland.

Gatos, B., Papamarkos, N. and Chamzas, C. (1997). Using curvature features in a multiclassifier OCR system. *Engineering Applications of Artificial Intelligence*, 10(2), 213-224. doi: 10.1016/s0952-1976(97)00002-x.

Graphics Recognition. Recent Advances and Perspectives. In J. Lladós & Y.-B. Kwon (Eds.), (Vol. 3088, pp. 87-99): Springer Berlin / Heidelberg.

Grozea, C., Gehl, C. and Popescu, M. (2009). ENCOPLOT: Pairwise Sequence Matching in Linear Time Applied to Plagiarism Detection. Paper presented at the *SEPLN*, Donostia, Spain.

Hassan, M. M. and Khatib, W. A. (2007). Similarity Searching In Statistical Figures Based On Extracted Meta Data. Paper presented at the *Proceedings of the Computer Graphics, Imaging and Visualisation*.

Huang, W. and Tan, C. L. (2007). A system for understanding imaged infographics and its applications. Paper presented at the *Proceedings of the 2007 ACM symposium on Document engineering*, Winnipeg, Manitoba, Canada.

Huang, W., Tan, C. and Leow, W. (2004). Model-Based Chart Image Recognition Graphics Recognition Algorithms and Systems. In K. Tombre & A. Chhabra (Eds.), (Vol. 1389, pp. 163-174): Springer Berlin / Heidelberg.

Huang, W., Tan, C. and Leow, W. (2004). Model-Based Chart Image RecognitionGraphics Recognition. Recent Advances and Perspectives. In J. Lladós & Y.-B. Kwon (Eds.), (Vol. 3088, pp. 87-99): Springer Berlin / Heidelberg.

Huang, W., Zong, S. and Tan, C. L. (2007). Chart Image Classification Using Multiple-Instance Learning. Proceedings of the 2007 *Applications of Computer Vision, 2007. WACV '07. IEEE Workshop on*. 2007/feb., 27.

Kasprza, J., Brandejs, M. and Kˇripac, M. (2009). Finding Plagiarism by Evaluating Document Similarities. Paper presented at the *SEPLN*, Donostia, Spain.

Kataria, S., Browuer, W., Mitra, P. and Giles, C. L. (2008). Automatic extraction of data points and text blocks from 2-dimensional plots in digital documents. Proceedings of the 2008 *Proceedings of the 23rd national conference on Artificial intelligence - Volume 2*. 2008. 1169-1174.

Li, Y., McLean, D., B, Z., O'Shea, J. D. and Crockett, K. (2006). Sentence Similarity Based on Semantic Nets and Corpus Statistics. *IEEE*, 18, 1138-1150.

Mantas, J. (1986). An overview of character recognition methodologies. *Pattern Recognition*, 19(6), 425-430. doi: 10.1016/0031-3203(86)90040-3.

Meyer zu Eissen, S., Stein, B. and Kulig, M. (2007). Plagiarism Detection Without Reference Collections. In R. Decker & H. J. Lenz (Eds.), *Advances in Data Analysis* (pp. 359-366): Springer Berlin Heidelberg.

Navarro, G. (2001). A guided tour to approximate string matching. *ACM Comput. Surv.*, 33(1), 31-88. doi: 10.1145/375360.375365

Ottenstein, K. J. (1976). An algorithmic approach to the detection and prevention of plagiarism. *SIGCSE Bull.*, 8(4), 30-41. doi: 10.1145/382222.382462

P Majumder, M. M., B.B. Chaudhuri. (2002). N-gram: a language independent approach to IR and NLP Paper presented at the *International Conference for Universal Knowledge*, India

Parker, A. and Hamblen, J. O. (1989). Computer algorithms for plagiarism detection. *Education, IEEE Transactions on*, 32(2), 94-99. doi: 10.1109/13.28038

Pinto, D., Civera, J., Barr\, A., \#243, n-Cede\, \#242, et al. (2009). A statistical approach to crosslingual natural language tasks. *J. Algorithms*, 64(1), 51-60. doi: 10.1016/j.jalgor.2009.02.005

Potthast, M., Barrón-Cedeño, A., Stein, B. and Rosso, P. (2011). Cross-language plagiarism detection. *Language Resources and Evaluation*, 45(1), 45-62. doi: 10.1007/s10579-009-9114-z

Potthast, M., Stein, B. and Anderka, M. (2008). A Wikipedia-Based Multilingual Retrieval Model. In C. Macdonald, I. Ounis, V. Plachouras, I. Ruthven & R.

White (Eds.), *Advances in Information Retrieval* (Vol. 4956, pp. 522-530): Springer Berlin Heidelberg.

Savva, M., Kong, N., Chhajta, A., Fei-Fei, L., Agrawala, M. and Heer, J. (2011). ReVision: automated classification, analysis and redesign of chart images. Paper presented at the *Proceedings of the 24th annual ACM symposium on User interface software and technology*, Santa Barbara, California, USA.

Shcherbinin, V. and Butakov, S. (2009). Using Microsoft SQL Server platform for plagiarism detection. Paper presented at the *SEPLN*, Donostia, Spain.

Su, Z., Ahn, B.-R., Eom, K.-y., Kang, M.-k., Kim, J.-P. and Kim, M.-K. (2008). Plagiarism Detection Using the Levenshtein Distance and Smith-Waterman Algorithm Paper presented at the *IEEE*.

Wang, Y., Phillips, I. T. and Haralick, R. M. (2002). Table Detection via Probability Optimization. Proceedings of the 2002 *Proceedings of the 5th International Workshop on Document Analysis Systems V*. 2002. 272-282.

Yang, L., Huang, W. and Tan, C. L. (2006). Semi-automatic ground truth generation for chart image recognition. Paper presented at the *Proceedings of the 7th international conference on Document Analysis Systems*, Nelson, New Zealand.

Yerra, R. and Ng, Y.-K. (2005). A sentence-based copy detection approach for web documents. Paper presented at the *Proceedings of the Second international conference on Fuzzy Systems and Knowledge Discovery - Volume Part I*, Changsha, China.

Yokokura, N. and Watanabe, T. (1998). Layout-based approach for extracting constructive elements of bar-charts Graphics Recognition Algorithms and Systems. In K. Tombre & A. Chhabra (Eds.), (Vol. 1389, pp. 163-174): Springer Berlin / Heidelberg.

Zhou, Y. and Tan, C. (2000). Hough technique for bar charts detection and recognition in document images. Proceedings of the 2000 *Image Processing, 2000. Proceedings. 2000 International Conference on DOI - 10.1109/ICIP.2000.899506*,

Zhou, Y. and Tan, C. L. (2001). Learning-based scientific chart recognition. Proceedings of the 2001 *4th IAPR International Workshop on Graphics Recognition, GREC2001*. 2001. 482-492.