INTERNATIONAL JOURNAL OF RESEARCH IN COMPUTER APPLICATIONS AND ROBOTICS
ISSN 2320-7345

A STUDY ON PLAGIARISM CHECKING WITH APPROPRIATE ALGORITHM IN DATAMINING

Hemalatha A.M[1], Ms. M. Subha, M.Sc.,M.Phil.,MCA., (PhD)[2]

[1] M.Phil Research Scholar /Kaamadhenu Arts & Science College/ hemasresearch@gmail.com
[2]Assistant Professor/Department of Computer Science/
Kaamadhenu Arts & Science College/subha.gmanoharan@gmail.com

ABSTRACT

The current plagiarism detection system was found to be too slow and takes too much time for checking. The matching algorithms are also dependent on the text's lexical structure rather than semantic structure. Therefore, it becomes difficult to detect the text paraphrased semantically. The big challenge is to provide plagiarism checking with appropriate algorithm in order to improve the percentage of finding result and time checking. The important question for the plagiarism detection problem in this study is whether it is possible to apply new techniques such as Semantic Role Labeling to handle plagiarism problems for text documents many documents are available on the internet and are easy to access. Due to this availability, users can easily create a new document by copying and pasting from these resources. Sometimes users can reword the plagiarized part by replacing the word with their synonyms. Motivation of the paper is to find the most plagiarism content that should be copied from anywhere identified in the efficient manner. Further it helps to as plagiarism detection process in applications to user or individual publish their journals.

Keywords: Plagiarism Checking, Algorithm, Data Mining & Text Mining

OBJECTIVE

Most empirical studies and analysis were undertaken by the academic community to deal with student plagiarism. In order to discriminate plagiarized documents from non-plagiarized documents, a correct selection of text features is a key aspect. The main objective of the paper is to find the more accurate plagiarism content in the documents with similar meaning and concepts are correctly identified in the efficient manner.

INTRODUCTION

Plagiarism defined as "unacknowledged copying of documents or programs". It can occur in many sectors. Plagiarism cases are an everyday topic, for example, in academics, journalism, and scientific research and even in politics. The recent case where the Hungarian President had to quit over a plagiarism scandal in April of 2012 is only one of the examples where copying and plagiarism can become a real problem. Most empirical studies and analysis were undertaken by the academic community to deal with student plagiarism. With the explosive growth of content found throughout the Web, people can find nearly everything they need for their written work, but detection of such cases can become a tedious task. For these reasons society needs to tackle this problem with computer-assisted approaches, and consequently, multiple studies in the field are being conducted.

In order to discriminate plagiarized documents from non-plagiarized documents, a correct selection of text features is a key aspect. There are many types of plagiarism mentioned by Hermann Maurer et al., such as copy and paste, redrafting or paraphrasing of the text, plagiarism of idea, and plagiarism through translation from one language to another. Nowadays, many documents are available on the internet and are easy to access. Due to this availability, users can easily create a new document by copying and pasting from these resources. Sometimes users can reword the plagiarized part by replacing the word with their synonyms. This kind of plagiarism is difficult to be detected by the traditional plagiarism detection system. Various methods can be implemented, ranging from document- comparison algorithms and systems to scan the Web, to approaches that utilize language-specific features, for example for the authorship-attribution task.

DATA MINING

The objective of data mining is to identify valid novel, potentially useful, and understandable correlations and patterns in existing data. Finding useful patterns in data is known by different names (including data mining) in different communities (e.g., knowledge extraction, information discovery, information harvesting, data archaeology, and data pattern processing).The term "data mining" is primarily used by statisticians, database researchers, and the MIS and business communities. The term Knowledge Discovery in Databases (KDD) is generally used to refer to the overall process of discovering useful knowledge from data, where data mining is a particular step in this process.

The additional steps in the KDD process, such as data preparation, data selection, data cleaning, and proper interpretation of the results of the data mining process, ensure that useful knowledge is derived from the data . Data mining is an extension of traditional data analysis and statistical approaches in that it incorporates analytical techniques drawn from a range of disciplines including, but not limited to,

- Numerical analysis,
- Pattern matching and areas of artificial intelligence such as machine learning,
- Neural networks and genetic algorithms.

While many data mining tasks follow a traditional, hypothesis-driven data analysis approach, it is commonplace to employ an opportunistic, data driven approach that encourages the pattern detection algorithms to find useful trends, patterns, and relationships. Essentially, the two types of data mining approaches differ in whether they seek to build models or to find patterns. The first approach, concerned with building models is, apart from the problems inherent from the large sizes of the data sets, similar to conventional exploratory statistical methods.

The objective is to produce an overall summary of a set of data to identify and describe the main features of the shape of the distribution. Examples of such models include a cluster analysis partition of a set of data, a regression model for prediction, and a tree-based classification rule. In model building, a distinction is sometimes

made between empirical and mechanistic models. The former (also sometimes called operational) seeks to model relationships without basing them on any underlying theory. The latter (sometimes called substantive or phenomenological) are based on some theory or mechanism for the underlying data generating process. Data mining, almost by definition, is primarily concerned with the operational.

The second type of data mining approach, pattern detection, seeks to identify small (but none the less possibly important) departures from the norm, to detect unusual patterns of behaviour. Examples include unusual spending patterns in credit card usage (for fraud detection), sporadic waveforms in EEG traces, and objects with patterns of characteristics unlike others. It is this class of strategies that led to the notion of data mining as seeking "nuggets" of information among the mass of data. In general, business databases pose a unique problem for pattern extraction because of their complexity. Complexity arises from anomalies such as discontinuity, noise, ambiguity, and incompleteness. And while most data mining algorithms are able to separate the effects of such irrelevant attributes in determining the actual pattern, the predictive power of the mining algorithms may decrease as the number of these anomalies increase.

MINING METHODOLOGY

- Mining different kinds of knowledge in databases**.** - The need of different users is not the same. And Different user may be in interested in different kind of knowledge. Therefore it is necessary for data mining to cover broad range of knowledge discovery task.

- Interactive mining of knowledge at multiple levels of abstraction**.** - The data mining process needs to be interactive because it allows users to focus the search for patterns, providing and refining data mining requests based on returned results.

- Incorporation of background knowledge. - To guide discovery process and to express the discovered patterns, the background knowledge can be used. Background knowledge may be used to express the discovered patterns not only in concise terms but at multiple level of abstraction.

- Data mining query languages and ad hoc data mining. - Data Mining Query language that allows the user to describe ad hoc mining tasks should be integrated with a data warehouse query language and optimized for efficient and flexible data mining.

- Presentation and visualization of data mining results. - Once the patterns are discovered it needs to be expressed in high level languages, visual representations. These representations should be easily understandable by the users.

- Handling noisy or incomplete data. - The data cleaning methods are required that can handle the noise, incomplete objects while mining the data regularities. If data cleaning methods are not there then the accuracy of the discovered patterns will be poor.

- Pattern evaluation. - It refers to interestingness of the problem. The patterns discovered should be interesting because either they represent common knowledge or lack novelty.

TEXT MINING

Text mining is the analysis of data contained in natural language text, which is sometimes referred to "text analytics", is one way to make qualitative or "unstructured" data usable by a computer. In other words, text mining is the discovery by computer of new, previously unknown information, by automatically extracting information from a usually large amount of different unstructured textual resources.

Qualitative data is descriptive data that cannot be measured in numbers and often includes qualities of appearance like color, texture, and textual description. Quantitative data is numerical, structured data that can be measured. However, there is often slippage between qualitative and quantitative categories. For example, a photograph might traditionally be considered "qualitative data" but when you break it down to the level of pixels, which can be measured.

The main motivation of our work is to study different existing tools and techniques of Text Mining for Information Retrieval (IR). Search engine is the most well known Information Retrieval tool. Application of Text Mining techniques to Information Retrieval can improve the precision of retrieval systems by filtering relevant documents for the given search query. Electronic information on Web is a useful resource for users to obtain a variety of information. The process of manually compiling text pages according to a user's needs and preferences and into actionable reports is very labor intensive, and is greatly amplified when it needs to be updated frequently.

Updates to what has been collected often require a repeated searching, filtering previously retrieved text web pages and re-organizing them. To harness this information, various search engines and Text Mining techniques have been developed to gather and organize the web pages. Retrieving relevant text pages on a topic from a large page collection is a challenging task.

Given below are some issues identified in Information Retrieval process: Traditional Information Retrieval techniques become inadequate to handle large text databases containing high volume of text documents. To search relevant documents from the large document collection, a vocabulary is used which map each term given in the search query to the address of the corresponding inverted file; the inverted files are then read from the disk; and are merged, taking the intersection of the sets of documents for AND, OR, NOT operations. To support retrieval process, inverted file require several additional structures such as document frequency of each lexicon in the vocabulary, term frequency of each term in the document.

The principal cost of searching process are the space requirement in memory to hold inverted file entries, and the time spend to process large size inverted files maintaining record of each document of the corpus as they are potential answers. Many terms in the query means more disk accesses into the inverted file, and more time spent to merge the obtained lists.

Presently, while doing query based searching, search engines return a set of web pages containing both relevant and non relevant pages, sometimes showing non relevant pages assigned higher rank score. These search engines use one of the following approaches to organize, search and analyze information on the web. In the first approach, ranking algorithm uses term frequency to select the terms of the page, for indexing a web page (after filtering out common or meaningless words). In the second approach structure of links appearing between pages is considered to identify pages that are often referenced by other pages.

Analyzing the density, direction and clustering of links, such method is capable of identifying the pages that are likely to contain valuable information. Another approach analyzes the content of the pages linked to or from the page of interest. They analyze the similarity of the word usage at different link distance from the page of interest and demonstrate that structure of words used by the linked pages enables more efficient indexing and searching.

Anchor text of a hyperlink is considered to describe its target page and so target pages can be replaced by their corresponding anchor text. But the nature of the Web search environment is such that the retrieval approaches based on single sources of evidence, suffer from weaknesses that can hurt the retrieval performance. For example, content-based Information Retrieval approach does not consider link information of the page while ranking the target page and hence affect the quality of web documents, while link-based approaches can suffer from incomplete or noisy link topology. The inadequacy of singular Web Information Retrieval approaches make a strong argument for combining multiple sources of evidence as a potentially advantageous retrieval strategy for Web Information Retrieval.

Text Mining, also known as knowledge discovery from text, and document information mining, refers to the process of extracting interesting patterns from very large text corpus for the purposes of discovering knowledge. It is an interdisciplinary field involving Information Retrieval, Text Understanding, Information Extraction, Clustering, Categorization, Topic Tracking, Concept Linkage, Computational Linguistics, Visualization, Database Technology, Machine Learning, and Data Mining.

Content-based text selection techniques have been extensively evaluated in the context of Information Retrieval. Every approach to text selection has four basic components:

- Some technique for representing the documents
- Some technique for representing the information needed (i.e., profile construction)
- Some way of comparing the profiles with the document representation
- Some way of using the results of the comparison

Searching the web has played an important role in human life in the past couple of years. A user either searches for specific information or just browses topics which interest him/her. Typically, a user enters a query in natural language, or as a set of keywords, and a search engine answers with a set of documents which are relevant to the query. Then, the user needs to go through the documents to find the information that interests him. However, usually just some parts of the documents contain query-relevant information. Our approach follows what has been called a term-based strategy: find the most important information in the document(s) by identifying its main terms, and then extract from the document(s) the most important information (i.e., sentences) about these terms .Moreover, to reduce the dimensionality of the term space, we use the latent semantic analysis, which can cluster similar terms and sentences into ‘topics’ on the basis of their use in context. The sentences that contain the most important topics are then selected for the summary.

Text analytics software can help by transposing words and phrases in unstructured data into numerical values which can then be linked with structured data in a database and analyzed with traditional data mining techniques. With an iterative approach, an organization can successfully use text analytics to gain insight into content-specific values such as sentiment, emotion, intensity and relevance. Because text analytics technology is still considered to be an emerging technology, however, results and depth of analysis can vary wildly from vendor to vendor. Text mining involves the application of techniques from areas such as information retrieval, natural language processing, information extraction and data mining.

Text mining usually involves the process of structuring the input text (usually parsing, along with the addition of some derived linguistic features and the removal of others, and subsequent insertion into a database), deriving patterns within the structured data, and finally evaluation and interpretation of the output. 'High quality' in text mining usually refers to some combination of relevance, novelty, and interestingness. Typical text mining tasks include text categorization, text clustering, and concept/entity extraction, production of granular taxonomies, sentiment analysis, document summarization, and entity relation modeling.

The purpose of Text Mining is to process unstructured (textual) information, extract meaningful numeric indices from the text, and, thus, make the information contained in the text accessible to the various data mining (statistical and machine learning) algorithms. Information can be extracted to derive summaries for the words contained in the documents or to compute summaries for the documents based on the words contained in them. Text mining offers a solution to this problem by replacing or supplementing the human reader with automatic systems undeterred by the text explosion. It involves analyzing a large collection of documents to discover previously unknown information. The information might be relationships or patterns that are buried in the document collection and which would otherwise be extremely difficult, if not impossible, to discover. Text mining can be used to analyze natural language documents about any subject, although much of the interest at present is coming from the biological sciences.

Figure 1 depicts a generic process model for a text mining application. Starting with a collection of documents, a text mining tool would retrieve a particular document and pre-process it by checking format and character sets. Then it would go through a text analysis phase, sometimes repeating techniques until information is extracted. Three text analysis techniques are shown in the example, but many other combinations of techniques could be used depending on the goals of the organization. The resulting information can be placed in management information system, yielding an abundant amount of knowledge for the user of that system.

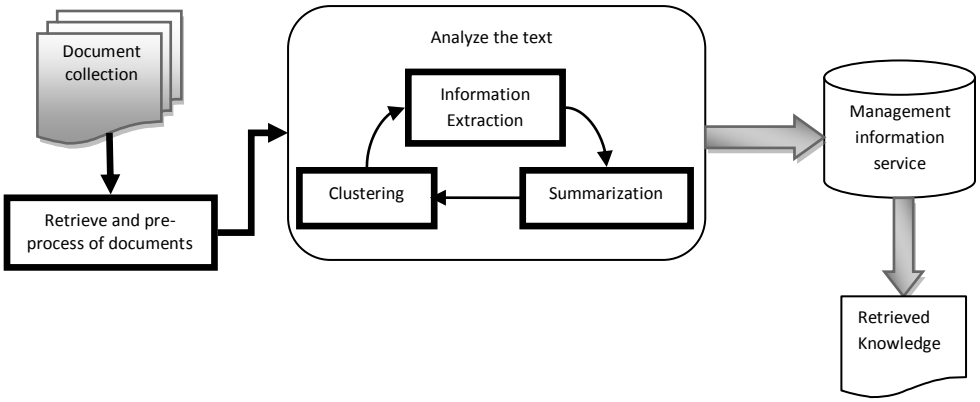


Figure 1: *An example of text mining*

A news topic is made up of a set of events and is discussed in a sequence of news stories. Most sentences of the news stories discuss one or more of the events in the topic. Some sentences are not germane to any of the events (and are probably entirely off-topic). Those sentences are called "off-event" sentences and contrast with "on-event" sentences. The order the stories are reported within the topic and the order of the sentences within each story combine to provide a total ordering on the sentences. We refer to that order as the natural order. The task of a system is to assign a score to every sentence that indicates the importance of that sentence in the summary: higher scores reflect more important sentences. This scoring yields a ranking on all sentences in the topic, including off- and on-event sentences. All sentences arriving in a specified time period can be considered together. They must each be assigned a score before the next set of sentences from the next time period. For this work, we have used a time period that has one story arriving at a time.

First, different kinds of plagiarism are organized into a taxonomy that is derived from a qualitative study and recent literatures about the plagiarism concept. The taxonomy is supported by various plagiarism patterns (i.e.,

examples) from available corpora for plagiarism. Second, different textual features are illustrated to represent text documents for the purpose of plagiarism detection. Third, methods of candidate retrieval and plagiarism detection are surveyed, and correlated with plagiarism types, which are listed in the taxonomy. 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 (CLIR), natural language processing, computational linguistics, artificial intelligence, and soft computing.

Text mining is also known as Text analytics, Knowledge Discovery from Text [KDT]. It is the process of extracting interesting and non-trivial patterns or structured text from unstructured documents. Text mining is defined as automatic discovery of previously Unknown information by extracting information from text. Data mining and Text mining are more similar in techniques. Data mining looks for patterns within structured data but Text mining looks for patterns with semi structured or unstructured data. The basic Text mining Techniques consists of

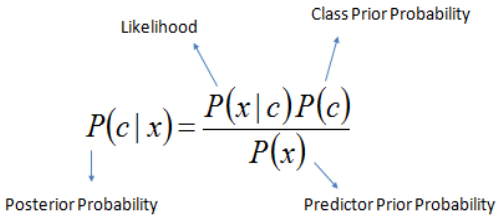
- Information Retrieval (IR), which collects and filters relevant document or information.
- Information Extraction (IE), which extracts useful information from the texts. IE deals with the extraction of particular entities and relationships.
- Data mining (DM), which extracts hidden, unknown patterns or information from data.

Text mining techniques can apply to structured or unstructured text

NAÏVE BAYESIAN CLASSIFIER

To overcome the problem of existing system, The proposed technique such as classification method. Naïve Bayesian is Simple ("naive") classification method based on Bayes rule. The Bayesian Classification represents a supervised learning method as well as a statistical method for classification. Assumes an underlying probabilistic model and it allows us to capture uncertainty about the model in a principled way by determining probabilities of the outcomes. It can solve diagnostic and predictive problems. Bayesian classification provides practical learning algorithms and prior knowledge and observed data can be combined. Bayesian Classification provides a useful perspective for understanding and evaluating many learning algorithms. It calculates explicit probabilities for hypothesis and it is robust to noise in input data.

The Naive Bayesian classifier is based on Bayes' theorem with independence assumptions between predictors. A Naive Bayesian model is easy to build, with no complicated iterative parameter estimation which makes it particularly useful for very large datasets. Despite its simplicity, the Naive Bayesian classifier often does surprisingly well and is widely used because it often outperforms more sophisticated classification methods. Bayes theorem provides a way of calculating the posterior probability, $P(c|x)$, from $P(c)$, $P(x)$, and $P(x|c)$. Naive Bayesclassifier assume that the effect of the value of a predictor (x) on a given class (c) is independent of the values of other predictors. This assumption is called class conditional independence.



$$P(c \mid \mathrm{X}) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

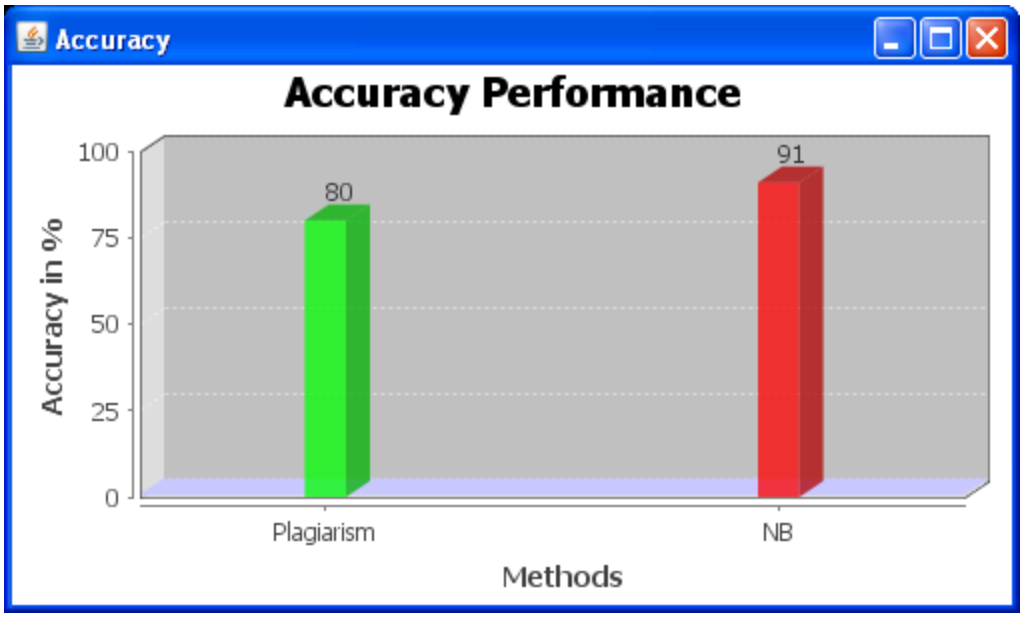
- P(c|x) is the posterior probability of class (target) given predictor (attribute).
- P(c) is the prior probability of class.
- P(x|c) is the likelihood which is the probability of predictor given class.
- P(x) is the prior probability of predictor.

Advantages

- In this approach, there use naïve Bayesian classification method; from this obtain more accuracy result than the existing system.
- Improved the performance of the system compared to existing approach
- If more than one author was written the document then this approach gives the accuracy result.

- **ACCURACY RATE**

- It defines as difference between existing and proposed systems result accuracy. Existing system or automated classification's accuracy having maximum of time for display result, minimum accuracy value and non realistic results. Proposed system or naïve Bayesian classifier of plagiarism detection system indicate that their performance depends on the type of plagiarism detection performance is increased and accurate result time is decreased than existing system. Intrinsic plagiarism detection using stylometry can overcome the boundaries of textual similarity to some extent by comparing linguistic similarity. Given that the stylistic differences between plagiarized and original segments are significant and can be identified reliably, stylometry can help in identifying disguised and paraphrased plagiarism.
- Stylometric comparisons are likely to fail in cases where segments are strongly paraphrased to the point where they more closely resemble the personal writing style of the plagiarist or if a text was compiled by multiple authors.



- *Fig 2 Accuracy comparison*

- This graph shows the accuracy rate of existing plagiarism detection and proposed NB based plagiarism detection based on two parameters of accuracy and methods such as existing and proposed system. From the graph we can see that, accuracy of the system is reduced somewhat in existing system than the proposed

system. From this graph we can say that the accuracy of proposed system is increased which will be the best one.

CONCLUSION

In this research explore the problem of text plagiarism and the possibility of its detection by the use of computer algorithms. In view of this, techniques and approaches to detect digital automated plagiarism detection have been introduced. One of the first problems the systems face is the collection of possible sources to compare the suspected documents with. This represent an entire problem in itself, and it is common that the ideal and real sources are not always available, limiting the potential of algorithms that compute similarity document-to-document. It was recently tested and studies utilizing different writing style markers are being introduced. In this research study a self-based information algorithm, whose basic idea is the use of a function to quantify the writing style based solely on the use of words.

But in this work, one important issue is if more than one author was written the document then the existing method will indicate as the plagiarized content. To overcome this problem, This proposed system introduce a classification method. Based on this classification approach we can obtain the accuracy result in the plagiarism detection

BIBLIOGRAPHY

[1] Kasprzak .J., & Brandejs .M, "*Improving the reliability on the plagiarism detection*

[2] *System*" – lab report for pan at clef 2010. In M. Braschler, D. Harman (2010).

[3] W.-j.L. Du Zou, Z. Ling, "*A Cluster-based Plagiarism Detection Method*", CLEF,

[4] (Notebook Papers/LABs/Workshops), 2010.

[5] Johnson and Wicheren, "*Extensive review of classical statistical algorithm*", (1998).

[6] Grieve, J. "Quantitative authorship attribution"," An evaluation of techniques",Literary and Linguistic Computing", pp. 22, 251–270, (2007).

[7] D.R. White, M.S. Joy, "*Sentence-based natural language plagiarism detection*", Journal of Education Resources in Computing 4 (4) (2004).

[8] Bao, J.-P., Shen, J.-Y., Liu, X.-D., Liu, H.-Y., & Zhang, X.-D. (2004)."*Semantic sequence Kin*", "*A method of document copy detection*", In H. Dai, R. Srikant, & C Zhang (Eds.), "Advances in knowledge discovery and data mining". Lecture notes in computer science, Vol. 3056, pp. 529–538.