Journal of Theoretical and Applied Information Technology
30th June 2018. Vol.96. No 12
© 2005 – ongoing JATIT & LLS
ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

AN ANALYSIS OF TEXT MINING FACTORS ENHANCING THE IDENTIFICATION OF RELEVANT STUDIES

MOUAYAD KHASHFEH, MOAMIN A. MAHMOUD, MOHD SHARIFUDDIN AHMAD

College of Computer Science and Information Technology, Universiti Tenaga Nasional

mou2ayad@gmail.com, moamin@uniten.edu.my, sharif@uniten.edu.my

ABSTRACT

The development of science and the spread of knowledge coincide with growing number of publications, and the volume of online content continue to grow at a rapid rate. For some submitted queries, the search engines may return thousands of documents of questionable relevancy. In this paper, we analyze the literature and identify the text mining factors that influence the identification of relevant studies. Five factors are identified which are Text Typography; Paragraph length; Term Frequency factor; Coordination; and Strict search. Subsequently, we propose an agent based-text mining model that facilitate the identification of relevant studies in big databases. The model consists of four components which are, interface, search process, parsing process, and storage. The interface provides a communication mean between a user and his/her counterpart agent (Personal Agent). In addition, it provides an input tool for user's search preferences. The second component is the search process that is operated by a pattern matching. The third process is the parsing that is operated by a text mining algorithm. The last part is the storage that is managed by Monitor Agent. The proposed framework would be useful in providing an alternative means of searching highly relevant studies from large databases.

Keywords: *Text Mining, Agent-based Model, Relevant Studies*

1.0 INTRODUCTION

Text mining facilitates knowledge discovery to extract useful hidden information from unstructured data such as textual data. Text mining also known as text analytics or text data mining, is the high-quality information discovering operation from unstructured textual data resources. Text mining techniques are used to solve some particular business issues. Text mining techniques are very important to organizations in order to analyse their wealth of textual information hence deriving valuable business visions to enhance the performance and increase the revenues.

Hotho et al. [1] stated that text mining is initially mentioned by Feldman and Dagan [2], who discussed the machine-supported analysis of text, by exploiting several techniques of information retrieval, information extraction and natural language processing (NLP) [1] [2]. However, the term text mining is coined from the concept of data mining.

Data mining involves the sifting of data records in databases to identify significant patterns that are useful for a decision-making process [3]. Among the data mining tasks such as classification or clustering, association rule mining is one particular task that extracts desirable information structures like correlations, frequent patterns, associations or causal between sets of items in transaction databases or other data stores [3] [4]. Association rule mining is widely used in many different fields like telecommunication networks, marketing, risk management, inventory control and others [3]. The association rule mining algorithm discovers association rules from a given database such that the rule satisfies a predefined value of support and confidence. The aim of using support and confidence thresholds is to ignore those rules that are not desirable, because the database is huge and users care about those frequently occurring patterns only [5].

The difference between data mining, association rule mining, and text mining is that in text mining, patterns are extracted from natural language text, but data mining and association rule mining patterns are extracted from databases [6]. According to Chiwara et al. [6], the text mining process steps are as follows:

- Text: This represents the given target document for mining which is in text format.
- Text processing: This step involves text clean up, format, tokenize and so on.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

- Text transformation (attribute generation): Generating attributes from the given processed text.
- Attribute selection: Selecting attributes for mining because not all generated attributes are suitable for mining.
- Data Mining (Pattern discovery): Mining the selected attributes and extracting desirable patterns.
- Interpretation and evaluation: It is about what is next, i.e. terminate, results well-suited for application at hand and so on.

Text mining is used by researchers to discover a particular desired information from unstructured and huge textual data [7]. Researchers applied their techniques in many domains such as, to discover the useful knowledge for enterprise decision, scientific papers to handle information retrieval [7], medical data set to extract useful knowledge [8], news articles for Stock Price Predictions [9]. The literature in this area mainly focus on developing text mining techniques [7] [8] [9] [10] [11], but less attention has been paid on developing a reasoning model that exploits text mining techniques to facilitate and accelerate extracting process [12] [13]. However, some researchers exploit multi-agent systems as a reasoning model for this need [4] [5].

In this paper, we develop a model that introduces the concept of agent-based text mining. The proposed model would be used by students and researchers to identify highly relevant papers from big data storage of papers. An agent gets a set of specific information from a user and by using text mining, it identifies the potential context of papers from the papers repository, and selects the relevant papers. The text mining could be used to enhance the efficiency of identifying relevant papers and the multi-agent system to speed up the process [14] [15] [16].

Our contribution in this paper is two-fold. Firstly, we analyze the literature on text mining and scientific papers 'styles to identify the potential mining factors of related studies. Secondly, we develop an agent-based text mining model that illustrates the process of identifying related studies utilizing the mining factors.

The next section dwells upon the related work on text mining and knowledge discovery. Section 3 identifies and analysis text mining factors. Section 4 illustrates the modeling of agent-based text mining. Section 5 concludes the paper.

2.0 RELATED WORK

2.1 Overview on knowledge discovery

Data growth on the Internet is increased dramatically due to the use of the internet by millions of people who publish articles, information, Statistical data etc. [17]. Thus, there is an urgent need to find methods to extract useful knowledge from this enormous amount of data. Extracting the knowledge from data requires using methodologies which in turn are classified under the name of Knowledge discovery [18].

Data resides in different forms [19], structured form (relational Database), unstructured form (text, documents, images) and semi-structured form (XML, JSON), these types [20] are used widely in big data analysis.

Some researchers assume that most of the available information found in structured form and saved in a relational database. Unfortunately, for many applications, available electronic information is in the form of unstructured documents rather than structured databases. Consequently, text mining is become an increasingly important aspect of knowledge discovery to discover useful knowledge from unstructured text [21].

Some people [22] don't know the difference between data mining and knowledge discovery, while others consider data mining a major stage of the Knowledge Discovery in Database (KDD) process. KDD refers to the overall process of extracting useful knowledge from data. Including the evaluation to make the decision of what qualifies as knowledge. Whereas Data mining refers to the application of algorithms for extracting patterns from data [23]. The KDD Process Steps Outline are as follows [22],

- Data Cleaning: Removing noise or outliers and inconsistent.
- Data Integration: Combining data from multiple data sources
- Data Selection: Retrieving Data relevant to the analysis stage from Database
- Data Transformation: Transforming or consolidating data into forms appropriate for mining by performing summary or aggregation operations.
- Data Mining: Applying intelligent methods to extract data patterns.
- Pattern Evaluation: Evaluating data patterns.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

- Knowledge Presentation: Representing knowledge.

2.2 Data types depending on Structuring

Structured Data: Structured data refers to information with a high degree of organization [24] [25] [26], the data in structured form resides in a fixed field within a record or file [25] such as relational databases and spreadsheets, storing data in this way is easily detectable and searchable using search operations and algorithms. Therefore, it will be relatively easy to enter, store, retrieve, and analyze at one time, and it is considered .

Structured data is basically based on data modeling. Data modeling [25] defines how data will be recorded and how they will be stored, processed and accessed, this requires [26] defining the field name and data type precisely such as (numeric, Date, alphabetic, currency) and any restrictions on the data input (number of characters; restricted to certain terms such as Mr., Ms. or Dr.; M; or F).

Structured data [25] has the advantage of being easily entered, stored, queried and analyzed. At one time, because of the high cost and performance limitations of storage, memory and processing, relational databases and spreadsheets using structured data were the only way to effectively manage data. Anything that could not fit into a tightly organized structure would have to be stored on paper in a filing cabinet.

Structured data is often managed using Structured Query Language (SQL) [7], SQL [27] is a special-purpose programming language created for designing, managing and querying data in relational database management systems. Originally developed by IBM in 1970 [25], it gained popularity [28] when the American National Standards Institute (ANSI) adopted the first SQL standard in 1986, and later developed commercially by Relational Software such as Microsoft Corporation and Oracle Corporation [25] [27] [28].

The discovery of knowledge sources from structured resources like database and data warehouse is called “Data mining” [19] [29] [30]. Data Mining [31] [32] in general is the process of finding and analysing data from different perspectives then categorizing and summarizing it into useful information. Companies can use this information to learn more information about their clients to develop its marketing plans which can be benefit in increasing sales and revenue and cutting costs. Technically, data mining [31] is the process of discovering correlations and patterns among tens of fields in huge relational databases. Data mining parameters include [30]:

- Association - looking for patterns where one event is connected to another event
- Sequence or path analysis - looking for patterns where one event leads to another later event
- Classification - looking for new patterns (May result in a change in the way the data is organized but that's ok)
- Clustering - finding and visually documenting groups of facts not previously known
- Forecasting - discovering patterns in data that can lead to reasonable predictions about the future (This area of data mining is known as predictive analytics.)

Semi-Structured Data: Semi-Structured Data is a combination of structured and unstructured date [25]. It is not organized into specialized repository like relational databases or other forms of data tables [33] [34], but nonetheless the semi-structured data contains associated information which called Metadata. It is also known as self-describing structure because it used some tags or other markers to distinguish certain elements and define hierarchies of records and fields within the data.

In Semi-Structured data, we may find some entities belong to the same class are grouped together but contain some different attributes and the difference in ordering of the attributes is normal [34], and also semi-structured data is less constrained than databases, Therefore, it is considered "loosely structured" [35].

For example, Word document file is considered unstructured date, but by adding some metadata tags as keywords which represent the document content and make it easier to be found when people search for those terms, then we can consider it a semi-structured data [33]. There is some types of semi-Structured data like XML and JSON (JavaScript Object Notation).

XML [36] is a good example of Semi-Structured data, there are no restrictions on the tags or nesting relationships. No required schema, where XML data is self-describing, structure and data are intertwined in one format. Because of the massive and rapid development of data, such as data on the Web.

XML gives users the freedom to change their data without constantly updating an associated schema. In

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

other situations, for data whose structure changes less often, XML optionally supports Document Type Definitions (DTDs) for restricting the tags and nesting rules. In either case, XML is ideal for exposing and exchanging a simple and convenient view of data.

JSON [37] (JavaScript Object Notation) is a lightweight data-interchange format. It is easy to read, easy to write for humans, and easy to parse and generate for machines. It is based on a subset of the JavaScript Programming Language. JSON [37] is a text format that is completely language independent but uses conventions that are familiar to programmers of all programming languages including Java, JavaScript, Python, C, C++, C#, Perl, and others. These properties make JSON an ideal data-interchange language.

Unstructured Data: Refers to information that either does not have a pre-defined data model and/or is not organized in a predefined manner [38]. It is more like human language. It doesn't fit nicely into relational databases like SQL, and searching it based on the old algorithms ranges from difficult to completely impossible. Examples include emails, text documents (Word docs, PDFs, etc.), social media posts, videos, audio files, and images [18].

Seth Grimes [39], a leading industry analyst on the confluence of structured and unstructured data sources, published an article that stated, "80% of business-relevant information originates in unstructured form, primarily text.

The discovery of knowledge sources that contain text or unstructured information is called "text mining" [19], Text mining tools could be technologies are capable of answering sophisticated questions and performing text search with an element of intelligence. A text mining application uses unstructured textual information and examines it in attempt to discover structure and implicit meanings hidden within texts [40]. Compared with the kind of data stored in databases, text is unstructured, amorphous, and difficult to deal with. Nevertheless, texts remain the most common vehicle for the formal exchange of information. The motivation for trying to extract information from it is compelling-even if success is only partial [41].

2.3 Multi-Agent Systems (MAS)

MAS has been hailed as a new paradigm for conceptualizing, designing, and implementing complex software systems. Agents are sophisticated computer programs that act autonomously on behalf of their users across open and distributed environments to solve complex computing problems [42] [43] [44]. Sycara [45] defined MAS as a loosely coupled network of problem solvers that interact to solve problems that are beyond the individual capabilities or knowledge of each problem solver. Technically MAS provides characteristics of having advantages of computational efficiency, reliability, extensibility, maintainability, robustness, maintainability and responsiveness, flexibility, and reuse [45] [51].

MAS is a system that contains two or more agents [46], at least one is an autonomous agent [47]. The agent is a software module that has capabilities to interact with each other and perform tasks independently. Interaction between agents can be conducted through an agent communication language (ACL) [Moamin]. MAS have been applied for various applications, including natural language processing [45], and data mining [48].

3.0 ANALYSIS OF FACTORS

In order to parse a document and extract keywords and terms along with their weights relevancy, a certain text mining algorithm need to be used. to analyses documents contents and explore structured data (tables and rows) from unstructured data (Documents).

Any text mining algorithm relies on some specific factors to analyze textual contents and these factors change as per the type of the contents, for example: HTML pages which consists of a URL and html tags as shown in table 1.

Table 1: Some HTML Tags Example

Tag	Description
<head>	Is a container for metadata (data about data), contains data about the HTML document, and defines the document title, character set, styles and other meta information.
<Body>	Defines the document's body containing all the contents of an HTML page, such as text, hyperlinks…. etc
<img>	To add an image inside the HTML page.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

<a href>	To add a link inside the HTML page.
<b>	Represents bold text in the HTML page
<title>	Defines the title of the HTML document defines a title in the browser toolbar
<h1><h2>……<h6>	To define the headers in the HTML page <h1> the most important heading. <h6> the least important heading.

As shown if Figure 1, these tags are very important to analyze a page and search keywords inside HTML, Search engine optimizations (SEOs) rely on the HTML tags to extract keywords, rank the page, and identify the page context.

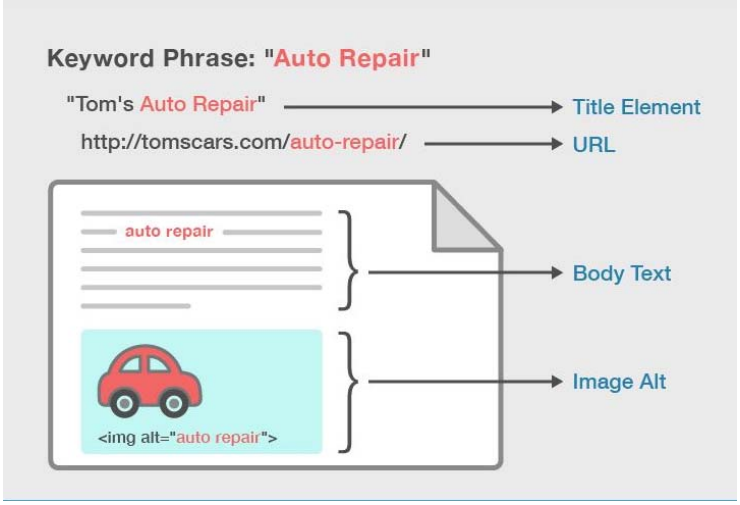

Figure 1: Using tags to analyze a page

Text mining relies on some factors to extract the keywords and identify the weight or the strength of each of them within the document [49][50]. As we mentioned the scope of this work is scientific publications, and the majority of scientific publications are in E-Document format like PDF and Word files, which means we cannot rely on specific tags like what the SEOs do to analyze HTML documents.

3.1 Text Typography Factor - ttf (Font Size and Wight)

Typographic hierarchy is about creating different levels of importance through typeface choice and text arrangement. Used sparingly, different typeface weights (and proportions) can guide the reader through long or complicated documents. Think of the various typeface weights as graphic road signs: a few, well-placed, will help the reader navigate the content. Too many will distract and confuse.

The font is a very important factor in writing papers, an author could highlight some words by changing font features for some particular words and terms in order to distinguish these words from others, as well as draw readers' attention to these texts.

Bold typefaces and large font size are good methods to achieve that, usually we change the fonts for specific text to larger or bold case in order to tell readers this text is important, like headlines and a title. Thus, the words with the largest font in a publication are the most important, this factor should be taken in consideration while analyzing documents to give the words with larger fonts or bold, higher strength more than the other words in lower fonts as shown in Figure 2.

Journal of Theoretical and Applied Information Technology
30th June 2018. Vol.96. No 12
© 2005 – ongoing JATIT & LLS
ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

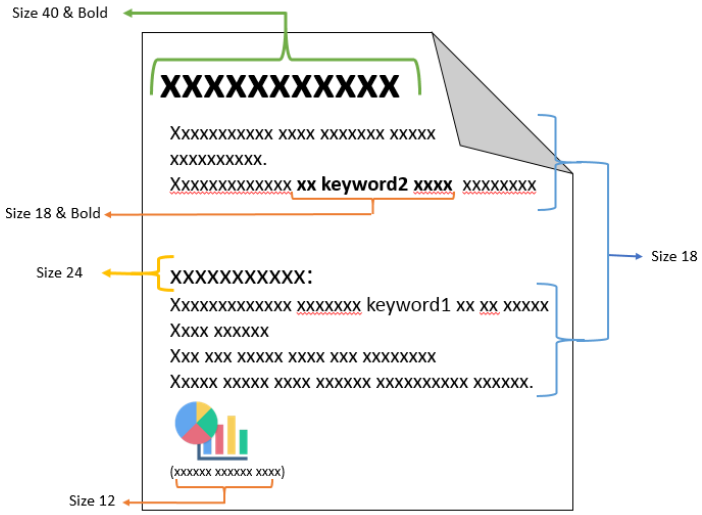

Figure 2 Font Features

Figure 2 consists of several font sizes (12, 18, 24, 40) and some of the words are bold, so in order to determine Text-Typography-Factor (ttf), a benchmark should be identified first to compare each font in document with it, hence all texts having a larger font than this benchmark would take higher score during parsing process for that document.

To determine the benchmark we divided the document words by the font size, and the size containing largest count of words would be the document's benchmark. Because it is the main font size used in this particular document and the score of this benchmark would be one. Then, any text smaller than benchmark would take the same score means one, and the larger font would take higher score which means these words are much more important. It is more probable that the words in this font identifies the context of the document more than the words in the benchmark font of this document.

Bold weights of type can easily establish priority. It gives the keyword higher score comparing with the normal words, usually the bold font is used to give the text higher priority, headings and titles are often set in bold type, bold type on a page adds a touch of graphic diversity to otherwise monotone pages of text copy.

3.2 Paragraph length (PL)

A keyword match found in a paragraph with a low count of total words would be more important than a keyword match found in a paragraph with a large number of keywords.

For example, if the keyword appears in a short text, such as the main title or sub-headings, it is more likely that the content of that text is about the keyword than if the same keyword appears in a much bigger body text.

Thus, if we have two documents, first document consist of 100 words and the second one contains 1000 words, the first document, which contains 100 words, would be more relevant than the second document containing 1000 words. Because the keyword in the first document was 1% of the text, while in the second document, it is only 0.1% of the text.

3.3 Term Frequency factor (TF)

TF means how often the keyword appears in a document, the more means higher weight. However, words may occur several time in a document, but might not be important and should be eliminated. For example, in English there are some words are used frequently like “the”,”are”,”is”,”for”… etc, these words called stop words, we can add all stop words to one list in order to removing them before starting the analyzing.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

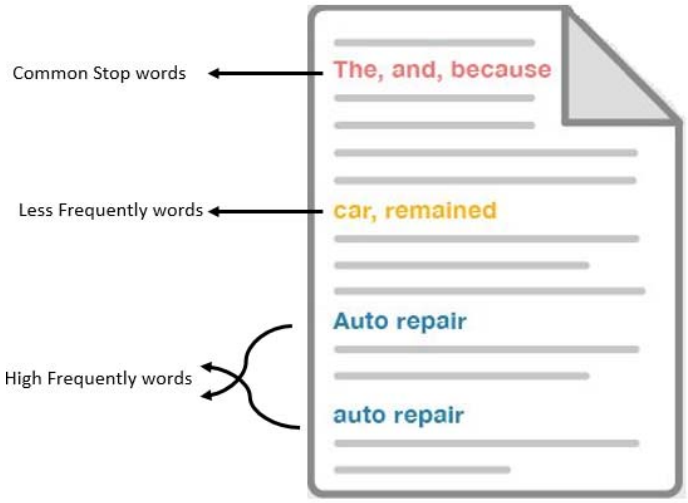

Figure 3 Term Frequency factor

3.4 Coordination (coord)

In case we have three keywords search request, then a document containing two keywords would be more relevant than another document containing just one keyword from the Query. Computing this value is very useful to order the results from more relevant to the less.

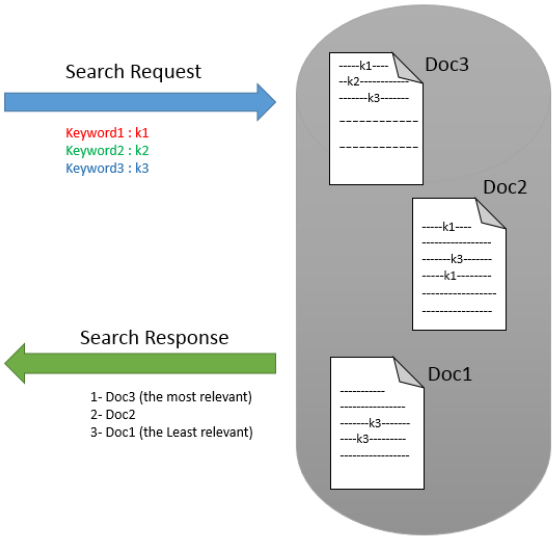

Figure 4 Coordination

3.5 Strict search not dynamic:

Comparing with the search engines like Google, we believe that for scientific publication we can return more relevant results if we make the searching process more specific. Google uses synonyms search to improve search process but for our scope, we plan to return exactly what a user/researcher wants. Online search engines may return extra results due to searching is made by using the Synonyms of the desired keyword, which can cause ambiguity during evaluating the results.

Journal of Theoretical and Applied Information Technology
30th June 2018. Vol.96. No 12
© 2005 – ongoing JATIT & LLS
ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

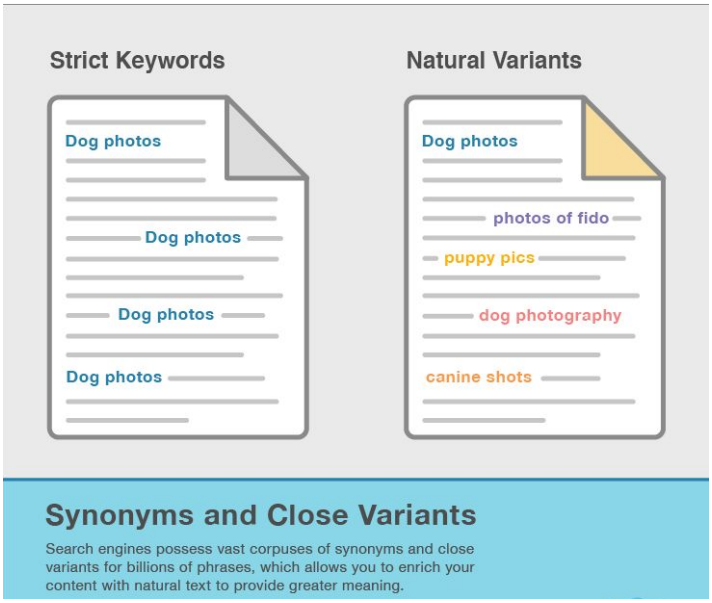

Figure 5 Strict search

4.0 MODELING OF AGENT-BASED TEXT MINING

4.1 Definitions

Definition 1: A keyword is any particular word or phrase used to characterize the document content. A keyword or combination of keywords can define the context of a document.

Definition 2: A keyword strength means how much can this keyword describes the idea of the document. In case a keyword strength is greater than the desired threshold value which provided by the end user, then the paper is considered meets the user's need.

Definition 3: A threshold relevancy is the lowest acceptable weight or strength of a document keyword.

Definition 4: An identity keyword that meets the maximum strength in a document. Identity keyword is used in order to measure the strength of the other keywords in the document.

Definition 5: A potential keyword is any keyword greater than or equal to the threshold relevancy value.

Definition 6: A weak keyword is any keyword less than the threshold relevancy value.

Definition 7: Documents Parser (as shown in Figure 6) is a the process of extracting the knowledge (structured data) from documents (unstructured data), the inputs of this process are documents and the outputs are keywords, thresholds as well as the terms from that particular document

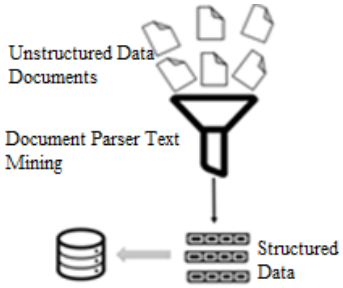

Figure 6: Document Parser

Definition 8: Storage is the main repository of the documents such as word files and PDF files, the storage could be FTP, local drive or cloud storage like Amazon storage (S3), Google Cloud storage (Bucket), Microsoft Azure storage (Block Blob) etc.

Definition 10: Global Knowledge Base (GKB) is a database contains the outputs of documents parsing, the text mining process extracts the keywords along with the thresholds relevancy for each keyword, and save these outputs inside GKB.

4.2 The Model

As shown in Figure 7, the proposed model consists of four components which are the interface, search process, parsing process, and the storage. The interface provides a communication mean between

Journal of Theoretical and Applied Information Technology
30th June 2018. Vol.96. No 12
© 2005 – ongoing JATIT & LLS
ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

the user and his/her counterpart agent (Personal Agent). In addition, it provides an input tool for users' search preferences. The second component is the search process that is operated by a pattern matching. The third process is parsing that is operated by a text mining. The last part is the storage that is managed by Monitor Agent.

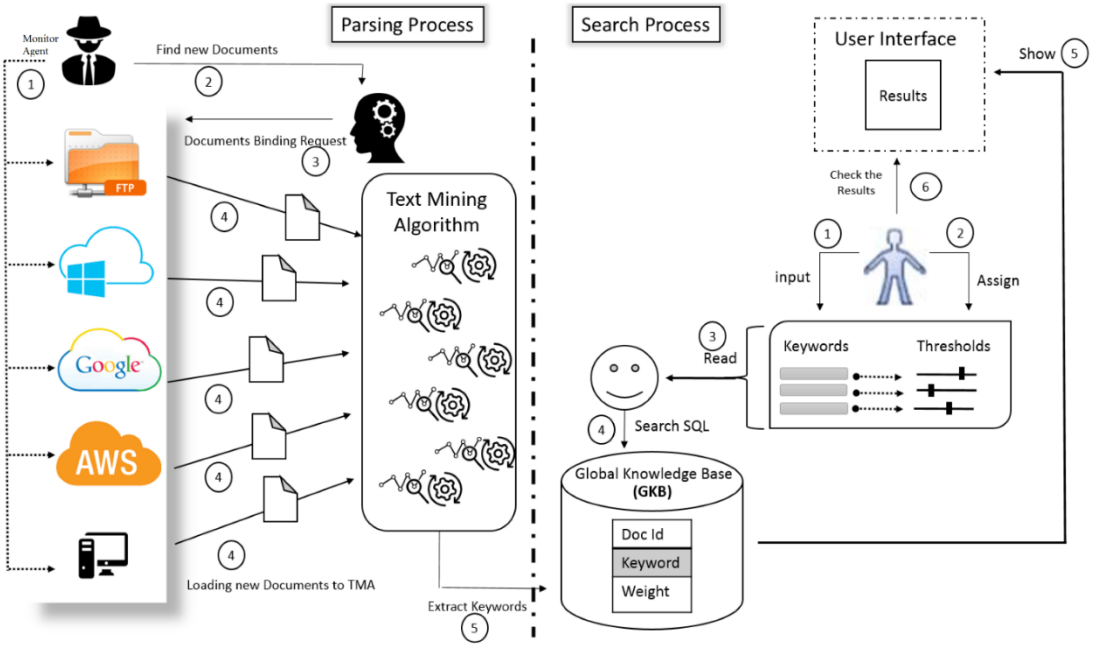

Figure 7: The proposed model

4.3 The Search Process

The process starts (1) by inserting desired keywords along with thresholds relevancy for each keyword. The Search agent collects (2) the inputs data and starts (3) searching inside the GKB database to find (4) the documents containing the desired keywords with threshold greater than or equal assigned thresholds. The agent returns (5) the results to the User interface with the possibility of saving the search for this user to send him/her further results later once a new document added and meets his search conditions. In this case, the system creates an agent on behalf of the user to provide him with these extra results; the generated agent would stay alive for a specific period could be specified by the user as well.

4.4 The Parsing Process

The process starts from documents repositories, once adding a new node to the module, the main agent of the parsing process assigns (1) a monitor agent to that storage node in order to catch any new added documents in that particular storage. The monitor agent then sends (2) a notification message contains document name and the event on this document to the main agent. The main agent in turn generates (3) a suitable number of agents to retrieve (4) and parse these documents in parallel. Parsing document can be achieved by applying a text mining process on each document to extract the keywords and the terms along with the thresholds relevancy for each keyword in the document. These data is saved (5) inside GKB database.

4.5 Personal Agent (PA) and Monitor Agent (MA)

As shown in Figure 8, when a user searches for documents using particular keywords with specific thresholds relevancy, he/she can enable the option "Further Results" as well, this option means that the user can get more results of the input parameters from the future documents, which would be added to the storage later. Hence, the system assigns the user's requests/inputs to the personal in order to start finding relevant documents on behalf of the user, and provide user with new results from the future documents accordingly. The user should also specify the period of the monitoring, for example for 10 months in this case the MA for this request would still active for 10 months and keep providing the user with further results for this period.

Journal of Theoretical and Applied Information Technology
30th June 2018. Vol.96. No 12
© 2005 – ongoing JATIT & LLS
ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

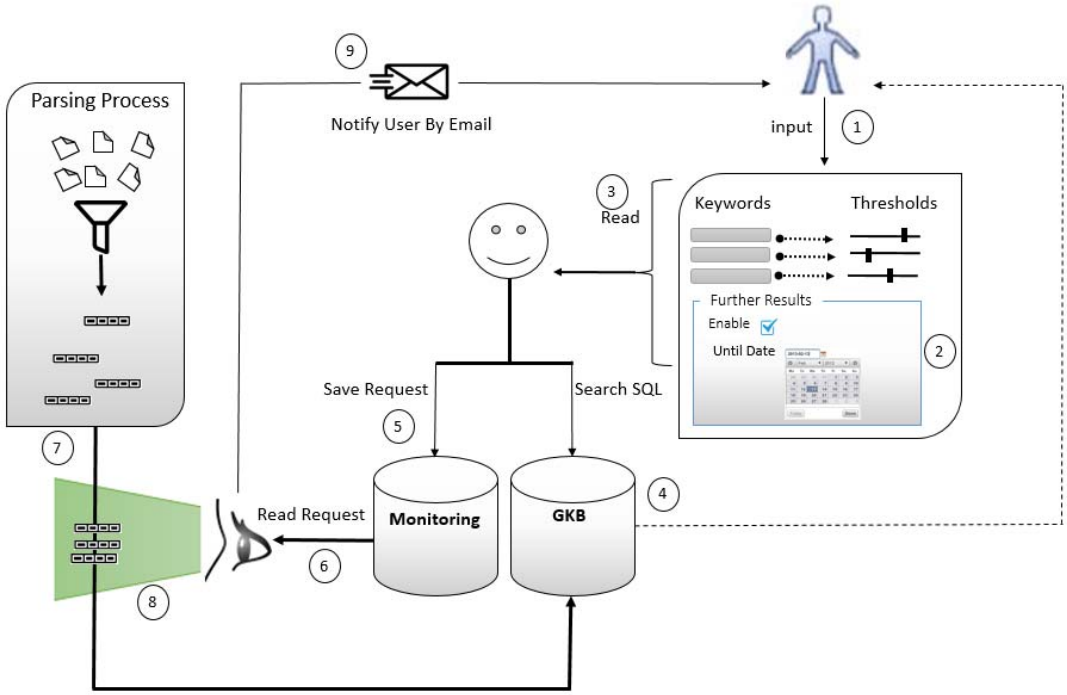

Figure 8: The process flow

After inserting the inputs (keywords, thresholds relevancy, Monitoring Date) the MA reads these inputs and search inside GKB and return the results to the user, and this agent would save the search request inputs inside another database called Monitoring database containing request inputs as well as requester email and monitoring validity period.

As shown in Figure 3, parsing process mines documents, extracts (7) keywords and their thresholds relevancy, and store these result inside GKB. During submitting the results to GKB, MA reads (6) the users requests from monitoring database and watches (8) parsing process results, if any results meets any User request, the PA would save this results and notify the requester later by email (9) that he/she has further results to his request.

5.0 CONCLUSION AND FURTHER WORK

The impact of the identified factors and the proposed model would be significant in providing universities; institutes; libraries, especially beginner, a new means in identifying highly relevant content from large databases with less time, cost, and effort, and limited skill. The identified factors are, Text Typography; Paragraph length; Term Frequency factor; Coordination; and Strict search. While the proposed model consists of four components which are, interface, search process, parsing process, and storage.

The limitation of this work is the proposed text mining model works very well with a certain articles' style that consists of title, abstract, keywords, and the body text. However, for other type of styles such as articles published online as essay or PowerPoint presentation, the accuracy of identifying relevant contents might drop.

In our future work, we shall develop a text mining algorithm using the identified factors. In addition, we shall develop a prototype that implements the proposed model. Using the prototype, we shall be able to study the efficiency, accuracy, and usability of the model in identifying highly relevant studies.

ACKNOWLEDGMENT

This project is sponsored by Universiti Tenaga Nasional (UNITEN) under the Bold Research Grant Scheme No. 10289176/B/9/2017/28.

REFERENCES

[1] Hotho, A., Nürnberger, A., & Paaß, G. (2005). A Brief Survey of Text Mining. LDV Forum 20(1), 19-26.

[2] Feldman R. and Dagan I. Kdt - knowledge discovery in texts. In Proc. of the First Int. Conf. on Knowledge Discovery (KDD), pages 112–117,1995.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[3] Kotsiantis S. & Kanellopoulos D., (2006). Association Rules Mining: A Recent Overview, International Transactions on Computer Science and Engineering, 32 (1), 71-82.

[4] Ogunde, A., Follorunso, O., Sodiiya, A., Oguntuase, J., & Ogunlleye G., (2011). Improved cost models for agent-based association rule mining in distributed database, Anale. Seria Informatica. IX (1), 231-250.

[5] Symeonidis A. L. & Mitkas P. A., (2006). Agent Intelligence through Data Mining, the 17th European Conference on Machine Learning and the 10th European Conference on Principles and Practice of Knowledge Discovery in Databases.

[6] Chiwara M., Al-Ayyoub M., Hossain M. S., Gupta R., Data Mining Concepts and Techniques Association Rule Mining, State University of New York, CSE 634, Chapter 8, 2006.

[7] Poelmans, J., Ignatov, D.I., Viaene, S., Dedene, G., Kuznetsov, S.: Text mining scientific papers: a survey on FCA-based information retrieval research. In: 12th Industrial Conference on Data Mining. LNCS, July 13-20, Berlin, Germany. Springer (2012).

[8] Liu, X. 2011. Learning from multi-view data: clustering algorithm and text mining application. Katholieke Universiteit Leuven, Leuven, Belgium.

[9] Aase K. Text Mining of News Articles for. Stock Price Predictions. Trondheim, June 2011. Master's thesis. Trondheim, 2011.

[10] Nahm U.Y., and Mooney R.J.. Text Mining with Information Extraction. In Proceedings of the AAAI 2002 Spring Symposium on Mining Answers from Texts and Knowledge Bases, pp. 60-67, Stanford, CA, March, 2002.

[11] N. Zhong, Y. Li, and S.T. Wu. Effective pattern discovery for text mining. IEEE Transactions on Knowledge and Data Engineering, 2011.

[12] Jusoh S. and Alfawareh H. M., "Agent-based knowledge mining architecture," in Proceedings of the 2009 International Conference on Computer Engineering and Applications, IACSIT. Manila, Phillipphines: World Academic Union, June 2009, pp. 602–606.

[13] Lai, K. K., Yu, L., & Wang, S. (2006). Multi-agent web text mining on the grid for enterprise decision support. In Advanced Web and Network Technologies, and Applications (pp. 540-544). Springer Berlin Heidelberg.

[14] Mostafa, S. A., Ahmad, M. S., Tang, A. Y., Ahmad, A., Annamalai, M., & Mustapha, A. (2014, April). Agent's autonomy adjustment via situation awareness. In Asian Conference on Intelligent Information and Database Systems(pp. 443-453). Springer, Cham.

[15] Mahmoud, M., Ahmad, M. S., & Yusoff, M. Z. M. (2016). Development and implementation of a technique for norms-adaptable agents in open multi-agent communities. Journal of Systems Science and Complexity, 29(6), 1519-1537.

[16] Mahmoud, M. A., Ahmad, M. S., Yusoff, M. Z. M., & Idrus, A. (2015). Automated multi-agent negotiation framework for the construction domain. In Distributed Computing and Artificial Intelligence, 12th International Conference (pp. 203-210). Springer, Cham.

[17] Lucas, W.E.N.D.Y. . (2000). The Noise Factor: Irrelevant Search Results on the World Wide Web. Information Resources Management Association, 354-358

[18] Ibmcom. (2016). Ibmcom. Retrieved 11 November, 2016, from

[19] Abiteboul,S.E.R.G.E. (1997). Querying semi-structured data. In N afrati, F.O.T.O. & G kolaitis, P.H.O.K.I.O.N. (Eds), Database Theory — ICDT '97 (pp. 1-18). Greece: Springer-Verlag Berlin Heidelberg.

[20] Schaefer , P.A.I.G.E. (2016). What's the Difference Between Structured and Unstructured Data?. Retrieved 12 November, 2016.

[21] J mooney, R.A.Y.M.O.N.D. & Bunescu, R.A.Z.V.A.N. . (2005). Mining Knowledge from Text Using Information Extraction. SIGKDD Explorations, 7(1), 3-10. Retrieved 22 November, 2016.

[22] Tutorialspointcom. (2016). Wwwtutorialspointcom. Retrieved 22 November, 2016.

[23] Hotho, A., Nürnberger, A., & Paaß, G. (2005). A Brief Survey of Text Mining. LDV Forum 20(1), 19-26.

[24] Brightplanetcom. (2012). BrightPlanet. Retrieved 22 October, 2016.

[25] Webopediacom. (2016). Webopediacom. Retrieved 12 November, 2016.

[26] Schaefer , P.A.I.G.E. (2016). What's the Difference Between Structured and

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Unstructured Data?. Retrieved 12 November, 2016.
[27] Wikipediaorg. (2016). Wikipediaorg. Retriev ed 12 November, 2016.
[28] Britannicacom. (2016). Encyclopedia Britannica. Retrieved 12 November, 2016.
[29] Fayyad, U.S.A.M.A. , Piatetsky-shapiro, G.R.E.G.O.R.Y. & Smyth , P.A.D.H.R.A.I.C. . (1997). From Data Mining to Knowledge Discovery in Databases. AAAI, 17(3), 37-54
[30] Techtargetcom. (2016). SearchSQLServer. Retrieved 20 November, 2016.
[31] Uclaedu. (2016). Uclaedu. Retrieved 20 November, 2016.
[32] Investopedia. (2003). Investopedia. Retrieved 20 November, 2016.
[33] Techtargetcom. (2016). WhatIscom. Retrieve d 14 November, 2016.
[34] Wikipediaorg. (2016). Wikipediaorg. Retriev ed 14 November, 2016.
[35] Upennedu. (2016). Upennedu. Retrieved 15 November, 2016.
[36] Stanfordedu. (2016). Stanfordedu. Retrieved 15 November, 2016.
[37] Jsonorg. (2016). Jsonorg. Retrieved 15 November, 2016.
[38] Libraryuunl. (2016). Libraryuunl. Retrieved 29 October, 2016.
[39] Skillsyouneedcom. Retrieved 29 October, 2016.
[40] H.Karanikas, C. Tjortjis and B. Theodoulidis, An approach to text mining using information extraction. In: Workshop of Knowledge Management: Theory and Applications in Principles of Data Mining and Knowledge Discovery 4th European Conference, (2000).
[41] I.Witten, H., Z. Bray, M. Mahoui and B. Teahan, Text mining: A new frontier for lossless compression. In: Proceedings of the Conference on Data Compression, 1999, pp. 198–207
[42] Mostafa, S. A., Ahmad, M. S., & Mustapha, A. (2017). Adjustable autonomy: a systematic literature review. Artificial Intelligence Review, 1-38.
[43] Mahmoud, M. A., Ahmad, M. S., & Yusoff, M. Z. M. (2016, March). A norm assimilation approach for multi-agent systems in heterogeneous communities. In Asian Conference on Intelligent Information and Database Systems (pp. 354-363). Springer, Berlin, Heidelberg.
[44] M.Aref, A multi-agent system for natural language understanding, International Conference on Integration of Knowledge Intensive Multi-Agent Systems (KIMAS '03), Boston, USA, 2003, pp. 36-40.
[45] K.Sycara,.P. Multiagent system, AI Magazine, Volume 19, No. 2, 1998, pp. 70-92 .
[46] Mahmoud, M. A., Ahmad, M. S., Yusoff, M. Z. M., & Mostafa, S. A. (2018, February). A Regulative Norms Mining Algorithm for Complex Adaptive System. In International Conference on Soft Computing and Data Mining (pp. 213-224). Springer, Cham.
[47] Mostafa, S. A., Gunasekaran, S. S., Ahmad, M. S., Ahmad, A., Annamalai, M., & Mustapha, A. (2014, June). Defining tasks and actions complexity-levels via their deliberation intensity measures in the layered adjustable autonomy model. In Intelligent Environments (IE), 2014 International Conference on (pp. 52-55). IEEE.
[48] W.Zhang and L. Zhang, A multiagent data warehousing (MADWH) and multiagent data mining (MADM) approach to brain modeling and neurofuzzy control. Inf. Sci. Inf. Comput. Sci. Vol. 167, No. 1-4, 2003, pp. 109-127.
[49] Gormley, C., & Tong, Z. (2015). Elasticsearch: The Definitive Guide: A Distributed Real-Time Search and Analytics Engine. " O'Reilly Media, Inc.".
[50] Segura-Bedmar, I., Carruana, A., & Martínez, P. (2016). LABDA at the 2016 BioASQ challenge task 4a: Semantic Indexing by using ElasticSearch. In Proceedings of the Fourth BioASQ workshop (pp. 16-22).
[51] Subramainan, L., Mahmoud, M. A., Ahmad, M. S., & Mohd Yusoff, M. Z. (2016). An emotion-based model for improving students' engagement using agent-based social simulator. International Journal on Advanced Science, Engineering and Information Technology, 6(6), 952-958.