A Text Mining Algorithm Optimising the Determination 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, sharif}@uniten.edu.my

***Abstract*—In this paper, we develop a text mining algorithm that influences the identification of relevant literature studies. The algorithm consists of three processes, detection process; preparation process; and mining process. The detection process includes the determination of document language and abstract and keywords. The Preparation includes the processes, split content to paragraphs; paragraph length determination; converting text to lower case; text typography factor; content tokenization, removing stop words. Finally, the mining includes the processes, regular expression; normalization; grouping and computing frequency. The proposed algorithm would be useful in providing an alternative means of searching highly relevant content from large databases.**

Keywords—Text Mining, Agent-based Model, Relevant Studies.

I. INTRODUCTION

The workflow of Text Mining contains some stages to reach target documents, these stages depend on techniques such as: Information retrieval (IR), Natural language processing (NLP), Information extraction (IE), Data mining (DM). However, by combining various stages together in correct order into a single workflow, we would achieve the Text mining workflow [1].

The literature in this area mainly focus on developing text mining techniques [2] [3] [4] [5] [6], but less attention has been paid on developing a reasoning model that exploits text mining techniques to facilitate and accelerate extracting process [7] [8]

However, some researchers exploit multi-agent systems as a reasoning model for this need [9] [10] [29] [30] [32]. In this paper, we develop a text mining algorithm that influences the identification of relevant literature studies. The algorithm is developed to be utilized by Multi-agent system. The algorithm consists of three processes, detection process; preparation process; and mining process. The detection process includes the determination of document language and abstract and keywords. The Preparation includes the processes, split content to paragraphs; paragraph length determination; converting text to lower case; text typography factor; content tokenization, removing stop words. Finally, the mining includes the processes, regular expression; normalization; grouping and computing frequency. The proposed algorithm would be useful in providing an alternative means of searching highly relevant content from large databases.

II. RELATED WORK

In the literature, three kinds of data are identified, Structured; semi-structured; and non-structured. Structured data refers to information with a high degree of organization [11, 12, 13], the data in structured form resides in a fixed field within a record or file [12] 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 modelling. Data modelling [12] defines how data will be recorded and how they will be stored, processed and accessed, this requires [13] defining in terms of fie ld name and data type precisely such as (numeric, Date, alphabetic, currency etc.) 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 [12] has the advantage of being easily entered, stored, queried and analysed. 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 couldn't 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) [12], SQL [14] is a special-purpose programming language created for designing, managing and querying data in relational database management systems. Originally developed by IBM in 1970 [12], it gained popularity 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 [12][14][15].

The discovery of knowledge sources from structured resources like database and data warehouse is called “Data mining” [16] [17] [18], Data Mining [19] [20] 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 [19] is the process of discovering correlations and patterns among tens of fields in

978-1-5386-7856-5/18/$31.00 ©2018 IEEE

huge relational databases. Data mining parameters include [18] [36]:

- 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 is a combination of structured and unstructured date [12]. It is not organized into specialized repository like relational databases or other forms of data tables

[18] [19], 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 [19], and also semi-structured data is less constrained than databases, Therefore, it is considered "loosely structured" [21]. 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 [18]. There is some types of semi-Structured data like XML and JSON (JavaScript Object Notation). XML [22] 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 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 [23] (JavaScript Object Notation) is a lightweight data-interchange format. It is easy to read and easy to write for humans. And also easy to parse and generate for machines. It is based on a subset of the JavaScript Programming Language. JSON [23] 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. Figure 2 shows an example of data represented in XML format and JSON format.

JSON Example

```
{"employees":[
    {"firstName":"John", "lastName":"Doe"},
    {"firstName":"Anna", "lastName":"Smith"},
    {"firstName":"Peter", "lastName":"Jones"}
]}
```

XML Example

```
<employees>
    <employee>
        <firstName>John</firstName> <lastName>Doe</lastName>
    </employee>
    <employee>
        <firstName>Anna</firstName> <lastName>Smith</lastName>
    </employee>
    <employee>
        <firstName>Peter</firstName> <lastName>Jones</lastName>
    </employee>
</employees>
```

Figure 2: XML format and JSON format [16]

Refers to information that either does not have a pre-defined data model and/or is not organized in a predefined manner [24]. 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 [25] Figure 3.

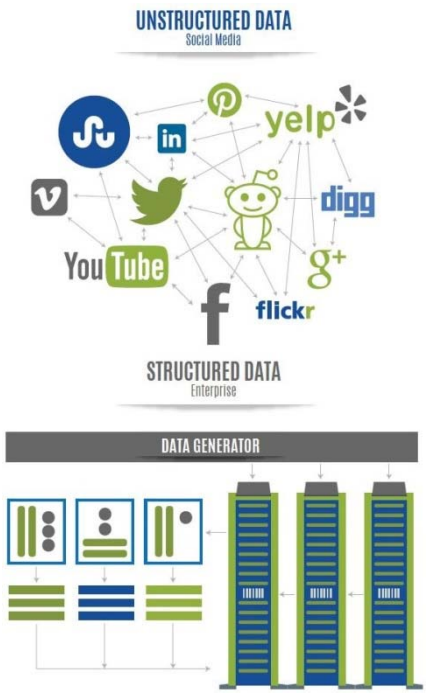

Figure 3 Structured Data [3]

Seth Grimes [26], 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." Figure 4.

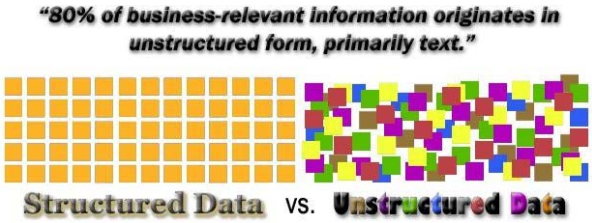

Figure 4 Structured VS. Unstructured Data [2]

The discovery of knowledge sources that contain text or unstructured information is called "text mining" [27], Text mining tools could be technologies are capable of answering sophisticated questions and performing text searches 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 the text [46]. 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 [28]. In the other hand, Multi agent systems 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 [31] [33] [34].

III. Text Mining Algorithm

In this Section, we present the developed algorithm to identify relevant studies from large databases that include thousands of articles. The algorithm consists of three stages, Detection Process; Preparation Process; and Mining Process. The detection process includes the determination of document language and abstract and keywords. The Preparation includes the processes, split content to paragraphs; paragraph length determination; converting text to lower case; text typography factor; content tokenization, removing stop words. Finally, the mining includes the processes, regular expression; normalization; grouping and computing frequency.

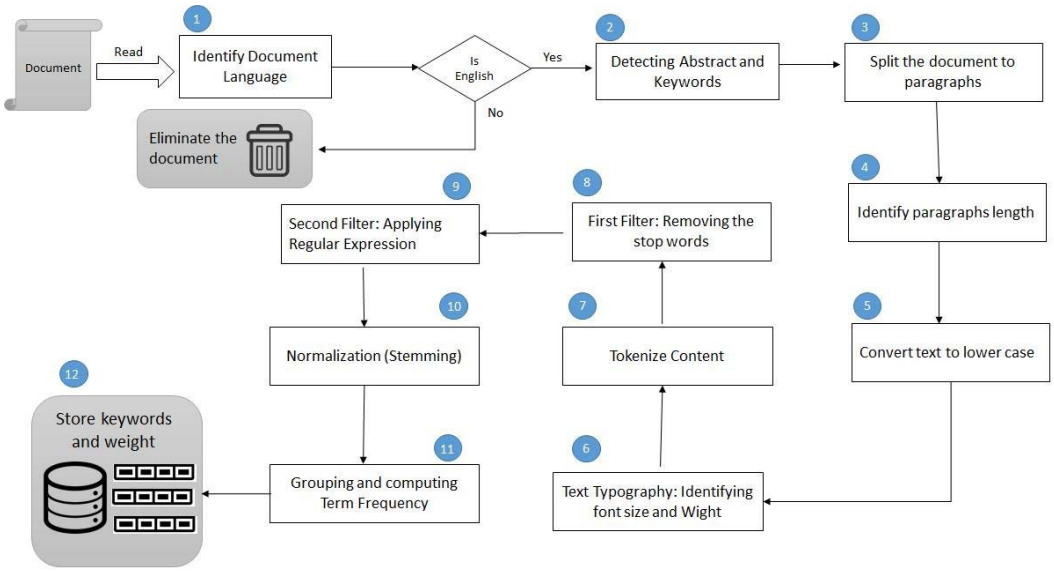

Figure 5 Text Mining Algorithm

A. Detection Process

Document Language: Prior parsing and analyzing documents, we need to verify whether a document has been written in English language and consequently eliminate non-English documents. Current practice uses ASCI code to identify the language, which we believe is not enough because there are some languages use English characters but the words are not English. In other words, using ASCI code, the text should match this RegEx "^[a-zA-Z0-9]*$" to be English text, this expression is used to match any letter from a-z in lower or upper case, as well as digits 0-9, and verify that no characters are allowed before or after this range. For example, the Arabic text will not pass this RegEx because the ASCI code for Arabic characters would be outside mentioned expression, but the Malaysian text would pass, like "Jalan" is not English word but it is written in English characters. Hence, using RegEx in not enough, we need to refer to a comprehensive dictionary contains many languages by comparing the words with other Languages. Such as Google Translator that provides a handy API to detect the language by sending an HTTP request using a specific URL.

Abstract and Keywords: An article's abstract and keywords is one of the major indicators of relevancy measurement and it is considered the most reflective part of the article content. Consequently, the existing words in abstract and keywords section would be scored higher than the words in other sections.

B. Preparation Process

Split the document to paragraphs: The next process is splitting the content to statements and paragraphs as shown in Figure 6, in order to start parsing process paragraph by paragraph, the words included in small paragraph would get higher score compared with the words within long paragraph. In this screenshot, we can see that the parsing process splits the document to list of paragraphs and returns 3 values Start index(From), End index(To) and the text of the paragraph. Splitting text to many paragraphs is mandatory to compute Paragraph length factor (PL).

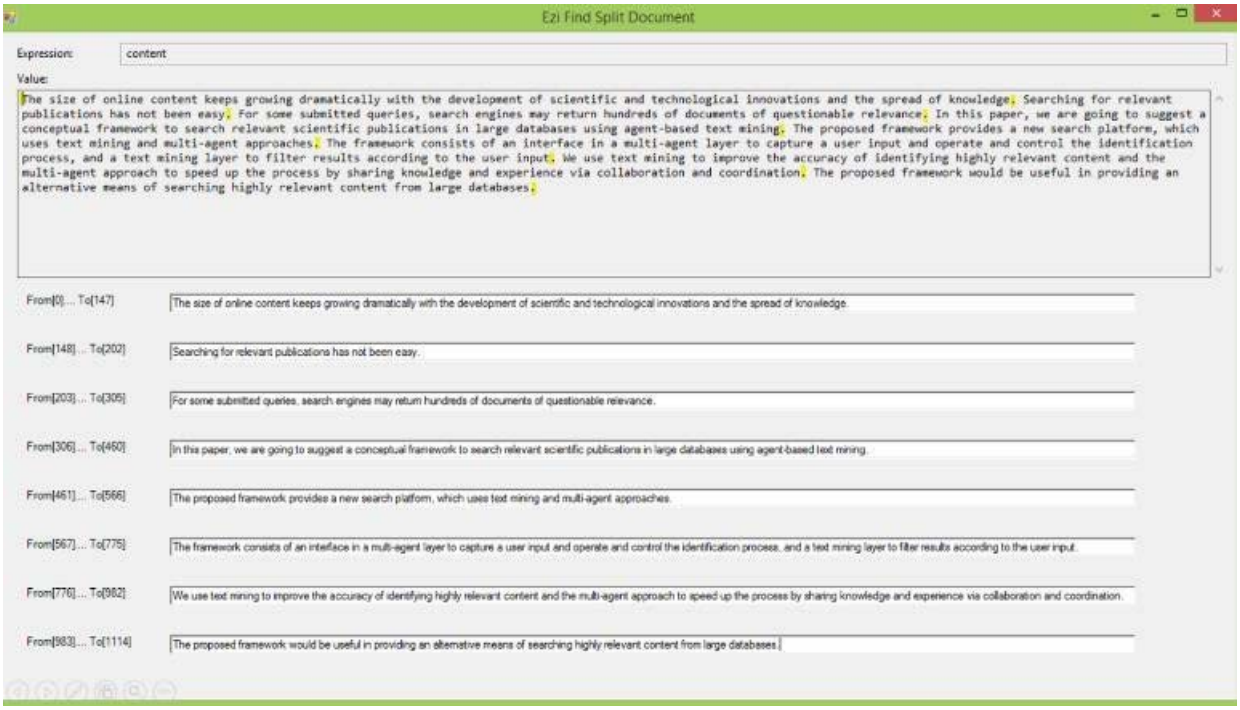

Figure 6 Split the document to paragraphs

Identify 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 a keyword appears in a short text, such as a 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.

Convert text to lower case: In general, all words in a document should be indexed in a standard format, means lower case or upper case, because the text mining algorithms should find the words and terms regardless of the case. Thus, in order to facilitate indexing document contents and increase the accuracy of the results, unification the characters' cases would be a need.

Text Typography Factor – ttf: Typographic hierarchy is used to create different levels of weight (importance) based on the typeface choice as well as the text order. Using different typeface weights is very useful to guide the reader through big and complicated articles, the writer can introduce the roadmap within the document so the reader can recognize the main points, titles and what the writer wants to highlight by playing around the font like using bold font or changing the font size. It very important to keep in mind that using it too many in a document, can cause distract and confuse. The font is a very important factor in writing publications, the publisher can 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 the reader 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 the reader this text is important,

like the headlines and the main title. Thus, the words with the largest font in a publication are the most important, so we decided to take this factor in our consideration while analyzing documents to give the words with larger fonts or bold, higher strength more than the other words in lower fonts.

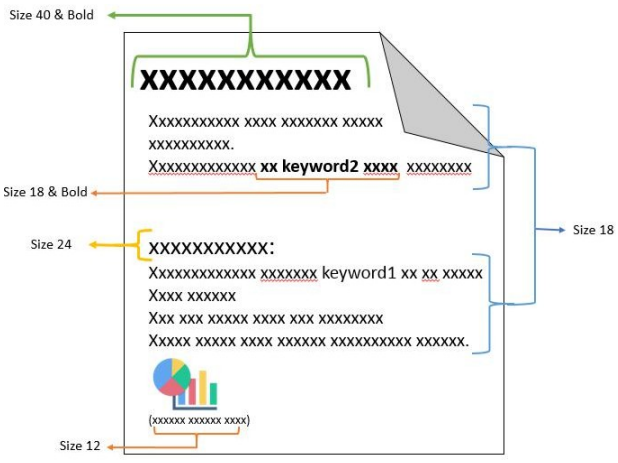

Figure 7 Text Typography Factor

Figure 7 consists of many font sizes (12, 18, 25, 40) and some of the words are bold, so in order to determine Text-Typography-Factor (ttf), 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, the document words is divided 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 1. Any text smaller than the benchmark would take the same score means 1, and the larger font would take higher score which means these words are more important, it is more likely that the words in the bigger font identifies the context of the document more than the words in the benchmark font. Bold weights of type can easily establish priority, it gives the keyword higher score comparing with the normal words. Generally, the bold font is used to give a text higher priority, headings and titles are often set in bold type.

Tokenize Content: Tokenization term refers to splitting text into many pieces such as words, phrases, symbols, keywords and other forms, and all of them called tokens. Tokens can be individual words, phrases or maybe full sentences, and some characters are ignored such as punctuation during tokenization process. The output of this process (tokens) is used as the input for another process like text mining and text analyzing. In computer science, Tokenization is considered a major process in the process of lexical analysis. In order to separate tokens, Tokenization process relies on some heuristics, by the following steps:

- Tokens are split by whitespace, punctuation marks or line breaks
- There is a possibility to include or exclude the whitespaces or the punctuation marks based on the need
- Tokens can be made up of all alphabetic characters, alphanumeric characters or just numeric characters.

Removing the stop words: Stop words are a set of usual words used in any language, English and non-English. The importance of stop words relies on the need of distinguishing them among the other words within a text and subsequently eliminate them before proceeding with the mining process as these words can cause wrong and confused results. Hence, removing these words from the text improve the accuracy of the results by focusing on the important words instead of including stop words, which are usually located with a high frequency within text. For example: if we pass this statement to a search engine "how to develop machine learning applications", let's say the search engine starts a search process to extract the web pages containing these terms "how", "to", "develop"," machine", "learning", "applications". in this case, the search engine will

return a lot of pages containing "how", "to" comparing with those containing indeed information about machine learning applications, due to using these words widely in English. Thus, if the mining algorithm can recognize and ignore these two terms, it can actually focus on returning results containing the desired keywords "develop"," machine", "learning", "applications".

C. Mining Process

Applying Regular Expression: Regular Expression is considered a very important and dynamic approach to express common patterns in a bunch of text strings. By changing or adding more regular expressions we can apply more filters and introduce more restrictions to eliminate unwanted words that negatively affect text parsing. Therefore, regular expression can be considered as information quality filters ensuring that a text string meets certain criteria. For example, by applying this RegEx "\b[a-zA-Z]{3,}\b" we can match words consisting of English chars only and the length of the word should be 3 chars or more, therefore eliminate words those containing numbers or non-English chars as well as eliminating words consisting of less than 3 chars.

Normalization (Stemming): In general, the underlying idea of stemming is identifying words having the same meaning but show up in different forms by eliminating suffixes and endings from the words and extract the root of the words, this kind of the identification is very important to figure out the weight of term correctly. Thus, increasing the effectiveness of information retrieval significantly. Stemming process is considered as pre-processing stage in text mining applications, it is a very popular requirement of NLP (Natural Language processing) functions, and actually, it is the most significant step in the Information Retrieval systems. The main aim of using stemming is reducing different grammatical forms of a word such as its adjective, noun, verb, adverb etc. to its root shape. In other words, reducing the inflectional forms and sometimes derivationally related forms of a word to a common base form. In text mining, the main purpose of stemming is improving retrieval effectiveness as well as reducing the indexing files size, means instead of indexing the one word in different shapes, we can index all words related to one meaning in just one shape.

There are many type of stemming algorithms such as Table lookup approach, Successor Variety, n-gram stemmers, Affix Removal Stemmers, Lovins, Paice, Porter. The Porter stemming algorithm is the most well-known one, it is a process for removing suffixes from the English words automatically. This can leverage especially in the field of information retrieval because in a typical environment a document is represented as a vector of words, or terms. Therefore, the terms having the same stem supposed to have a similar meaning. For example, this list of words related to one stem as shown in Figure 8.

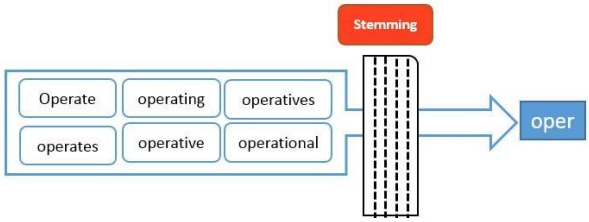

Figure 8 List of words related to one stem

Grouping and computing: Term Frequency factor(TF) TF means how often the keyword appears in the document. The more often means higher weight. The document containing seven mentions of the same keyword is much likely to be relevant than a document containing that keyword only once.

IV. Conclusion

This paper presents a combination between text mining and data mining with Multi agent system in order to extract potential context of scientific publication from particular scientific publication warehouse and thus define which publication are potentially relevant to a searcher's need. The proposed algorithm gives the ability to the user to input the keywords In addition assigning a threshold value for each one. The algorithm consists of three processes, detection process; preparation process; and mining process. The detection process includes the determination of document language and abstract and keywords. The Preparation includes the processes, split content to paragraphs; paragraph length determination; converting text to lower case; text typography factor; content tokenization, removing stop words. Finally, the mining includes the processes, regular expression; normalization; grouping and computing frequency. The proposed algorithm would be useful in providing an alternative means of searching highly relevant content from large databases.

In our future work, we shall formulate a scoring system for the developed algorithm and subsequently develop a prototype. Using the prototype, we shall be able to study the efficiency, accuracy, and usability of the model in identifying highly relevant studies.

References

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

[2] Poelmans, J., Ignatov, D.I., Viaene, S., Dedene, G., Kuznetsov, S.: Text

[3] mining scientific papers: a survey on FCA-based information retrieval

[4] research. In: 12th Industrial Conference on Data Mining. LNCS, July 13-20, Berlin, Germany. Springer (2012).

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

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

[7] 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.

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

[9] Jusoh S. and Alfawareh H. M., "Agent-based knowledge mining architecture," in Proceedings of the 2009 International Conference on

[10] Computer Engineering and Applications, IACSIT. Manila, Phillipphines: World Academic Union, June 2009, pp. 602–606.

[11] 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.

[12] 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.

[13] 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.

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

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

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

[17] Wikipediaorg. (2016). Wikipediaorg. Retrieved 12 November, 2016.

[18] [15] Britannicacom. (2016). Encyclopedia Britannica. Retrieved 12 November, 2016.

[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] 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

[21] Techtargetcom. (2016). SearchSQLServer. Retrieved 20 November, 2016.

[22] Uclaedu. (2016). Uclaedu. Retrieved 20 November, 2016.

[23] Investopedia. (2003). Investopedia. Retrieved 20 November, 2016.

[24] Upennedu. (2016). Upennedu. Retrieved 15 November, 2016.

[25] Stanfordedu. (2016). Stanfordedu. Retrieved 15 November, 2016.

[26] Jsonorg. (2016). Jsonorg. Retrieved 15 November, 2016.

[27] Libraryuunl. (2016). Libraryuunl. Retrieved 29 October, 2016.

[28] Ibmcom. (2016). Ibmcom. Retrieved 11 November, 2016.

[29] [26] Skillsyouneedcom 2011-2016. (2016). Skillsyouneedcom. Retrieved 29 October, 2016.

[30] 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.

[31] Witten, H., Z. Bray, M. Mahoui and B. Teahan, Text mining: A new

[32] frontier for lossless compression. In: Proceedings of the Conference on Data Compression, 1999, pp. 198–207.

[33] 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.

[34] 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.

[35] 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.

[36] 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.

[37] Mostafa, S. A., Ahmad, M. S., & Mustapha, A. (2017). Adjustable autonomy: a systematic literature review. Artificial Intelligence Review, 1-38.

[38] Mostafa, S. A., Ahmad, M. S., Tang, A. Y., Ahmad, A., Annamalai, M.,

[39] & 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.

[40] 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.

[41] Mahmoud, M. A., & Ahmad, M. S. (2016, August). A prototype for context identification of scientific papers via agent-based text mining. In Agent, Multi-Agent Systems and Robotics (ISAMSR), 2016 2nd International Symposium on (pp. 40-44). IEEE