Discovering Contextual Knowledge with Associated Information in Dimensional Structured Knowledge Bases

Johannes Zenkert, Alexander Holland and Madjid Fathi

University of Siegen
Institute of Knowledge Based Systems and Knowledge Management
Department of Electrical Engineering and Computer Science
Germany
johannes.zenkert@uni-siegen.de, alexander.holland@uni-siegen.de, madjid.fathi@uni-siegen.de

***Abstract*—The visualization and simplification of complex semantically-related knowledge is one of the main challenges in knowledge discovery. In this regard, the knowledge map is a good visualization instrument to represent and provide suitable information with analysis potential. Multidimensional knowledge bases aim to support this objective and store automatically extracted facts and their dimensional relations from textual knowledge resources. In this paper, a dynamic layout structure for knowledge maps based on dimensional information is introduced. The Concept of the Imitation of the Mental Ability of Word Association (CIMAWA) is applied in this approach to create a graphical structure as arrangement of associated information on different levels of textual information.**

***Keywords*—Knowledge Maps, Word Association, Integrate Knowledge**

I. Introduction

Knowledge maps are in wide-spread usage to represent knowledge in a graphical form. Especially in organizations, the knowledge map is a useful tool to visualize the company's experts, other knowledge holders, sources of knowledge and strategic position of the company [1]. Furthermore, knowledge maps are used for the access on knowledge in time, identification of knowledge assets, identification of knowledge flow, identification of existing knowledge resources, organizational restructuring, identification of knowledge gaps, team building and identification of untapped knowledge [2]. Similar representation tools are mind maps, concept maps and graphic organizers [3] which are not taken into consideration.

In organizations, business processes can be greatly improved and done more effectively and efficiently by the integration of relevant knowledge. In this regard, the creation of a knowledge source map is a very useful approach to recognize the available knowledge in the organization and to identify the lack of relevant knowledge in the knowledge identification phase of the organization's knowledge management [2][4]. Once experts and knowledge assets are identified and described, maps provide the current overview of accessible information. However, knowledge source maps are often considered as static constructs whereas the actual knowledge cannot be considered as a static organizational asset and is changing over the time. In dynamic (changing) enterprise structures, knowledge maps need to be updated frequently. The knowledge itself is highly depending on the employee's and expert's constantly varying tacit knowledge which makes it also difficult to extract and describe it a knowledge map.

Concept map, mind map, idea map, concept circle diagram, semantic map, cognitive map, process map, conceptual map, knowledge flow map, causal map, social mess map, ontology, petri net, cluster vee diagram, thesauri, visual thinking network, topic map and perceptual map have been identified as tools and techniques for knowledge map creation [2]. For each of the aforementioned map types different modified or specialized versions are existing. For example, the knowledge asset map is a type of conceptual map. This instrument is used to visualize how the available organizational knowledge can be accessed and more importantly where it is (physically) stored [5]. Knowledge asset maps are also very suitable to visualize the allocation of knowledge into distributed knowledge bases.

To overcome the limitations of static or inconsistent knowledge maps in any type, maps must be dynamic and automatically updated based on new information which is retrieved in the organization and stored in local or distributed knowledge bases. For textual resources, a dimensional structuring of the knowledge base is very beneficial. Different dimensional information which has been identified from the text with text mining methods can be stored in retrievable form in the knowledge base. In this regard, dimensional information can be derived from textual resources in the forms of meta data information and text analysis results (e.g. entity, sentiment, topic, associative information) [6].

For the automatic creation of knowledge maps, this paper suggests an approach which utilizes the Concept of the Imitation of the Mental Ability of Word Association (CIMAWA) [7] to describe the associative relationships between knowledge facts. By the combination of all calculated word association strengths, the numeric values of CIMAWA can be used as

978-1-5090-1897-0/16/$31.00 ©2016 IEEE

proximity measurements to describe how similar knowledge facts are and how close they are related to other knowledge facts from the knowledge base. For the visualization of associative knowledge maps, the CIMAWA calculations can be used for the arrangement of entities and their associated information.

Section 2 presents related work in the field of knowledge maps. The levels of associated knowledge based on document structures and named entities are described in Section 3. In Section 4 the application of the CIMAWA is mentioned as a method to discover related content based on word associations in the knowledge base. Furthermore, the calculation of CIMAWA is introduced as a proximity measure to create distance values between knowledge facts and the associative content for graphical visualization. Section 5 discusses the application of aforementioned theoretic background in dynamic knowledge maps and virtual reality applications. Section 6 shortly summarizes the results of this paper.

II. Related Work

For the creation of knowledge maps, data mining methods are commonly used and have been successfully applied in [8]. However, data mining methods cannot be directly applied on unstructured data in the form of text and require different preprocessing steps. Knowledge facts which can been discovered and extracted from textual resources with text mining methods are considerable for the integration into a knowledge map in order to retrieve and visualize them. Text mining methods have been already applied in [9] to generate a hierarchical knowledge map based on online Chinese news. Furthermore, associations have been used in text mining research to create knowledge maps [10].

The hybrid association measurement of the CIMAWA can be used to identify associative and contextual knowledge [7]. A high numeric value of word association strength which is calculated with CIMAWA methodology [7] often indicates a semantic relationship between the parts of the knowledge facts. Named Entities which can be found with text mining methods from the textual resources are considered as one of the dimensions in multidimensional knowledge bases [6]. In this way, persons, locations, organizations, brands, etc. are kept and represented with related information and facts within the knowledge base. By selecting different dimensions from the knowledge base, the entities together with all related information can be compared and illustrated in a knowledge map. Because of the flexibility in dimensional structure, knowledge maps can be created for specific use cases with desired entity-associated information.

Maps which have been created through multiple dimensional selections are able to provide associative content by the knowledge base through the aforementioned represented dimensional relationships and associations. Furthermore, knowledge maps that utilize extracted information from knowledge bases are dynamic in their nature because word associations are changing over the time by adding relevant facts or removing irrelevant information from the knowledge base. By considering these temporal changes, the dynamic knowledge map has large analysis potential. The dynamic knowledge map has been approached before. Woo et al. [11] suggested to use a dynamic knowledge map in the architecture, engineering and construction industry for tacit knowledge utilization.

Balaid et al. [2] analyzed the key challenges and barriers for knowledge maps in related literature. Reference [2] collected the main challenges from [12] as mapping of dynamic knowledge, cross-boundary knowledge mapping, knowledge representation, organizational leadership and culture and mapping of tacit knowledge. The mapping of complex structures to graphic symbols, the outdated map and the map usage of unauthorized users are mentioned in [4]. Moreover, the failure to understand the business process and the lack of understanding the knowledge flows are described in [13]. Dang et al. [14] are mentioning the interactive search, analysis and also the inclusion of essential document sources as key challenges.

In the following sections, some of the key challenges and barriers of knowledge maps within the literature are addressed and specific solutions for them are further elaborated. Section 3 is related to knowledge representation of organizational knowledge. Different levels of associated knowledge are discussed and the complexity of structures from knowledge items is reduced with text mining methods. Section 4 addresses the aforementioned document inclusion, the interactive search and the analysis within knowledge maps. Furthermore, the discussed frequent re-calculation of word association strengths provides a solution for the problem of outdated maps. Section 5 provides an approach for the dynamic knowledge map and discusses graphical representation of knowledge.

III. Levels of associated knowledge

Semantically-related information can be identified on different analysis levels. Textual resources in total or parts from the content can be considered as related if the same or semantically-related expressions are used in the text for e.g. customer names, product names or topics. Moreover, textual content is often also explicitly referencing to other sources of information. For organizational knowledge bases, the most common format to store information is in the form of (electronic) documents. Other formats like notes, e-mail communication or other external resources like web data are also used by employees to digitally access, create and store relevant knowledge. The different formats within an organizational knowledge management are considered as knowledge items. Documents can be organized by Document Management Systems (DMS) by indexing or keyword labeling. However, the actual information which is mentioned in any part of the document is hidden and not directly accessible.

For each specific format, different levels need to be considered in the analysis. While documents may discuss different topics in total, a paragraph within the document could be focused on one specific aspect. On smaller level, a sentence could indicate a fact that is maybe relevant to be extracted from the document and stored in the knowledge base. Furthermore,

the level of words inside a document could offer interesting insights to find semantically associated content for each word individually. Especially, after a sentence has been tokenized, disambiguated and analyzed with Named Entity Recognition (NER), the entity information can be directly used to create semantic relationships to other entity information within the knowledge base. As a result, all extracted, semantically-related content can be stored in the knowledge base by utilizing the dimensional structure.

A. Document Structures

The structure of documents needs to be subdivided in levels to capture the different association possibilities. On the highest level, the document level, different pieces of information within the whole document can be considered for contextual or associated knowledge. For example, topics can be identified through topic detection in the text mining process. The available methodologies in this area reach from keyword extraction to detection of multi-topic structures in documents [15].

Meta data from the document can be also collected and added to extracted information from the document in order to provide structural information and relationships. In this way, the meta data is used as dimensional information in multidimensional knowledge bases to retrieve the document-specific content via different document properties. The author of the document, title, number of pages, comments, category and timestamp information are few examples for the meta data information from a document which can leverage the provision of associated and contextual information.

On more detailed levels, like on a page level or on the paragraph level, the amount of information is greatly reduced and more precise associations for the content can be achieved. However, if the contextual information is reduced within a document, the ambiguousness of the remaining textual content increases. By looking further into a paragraph, the sentence level provides further meaningful information. Sentences in form of a statement can be directly integrated into the knowledge base as knowledge facts. On the sentence level, part-of-speech and link grammar help to identify and retrieve the knowledge facts which are stored in the knowledge base if they provide additional insight. A sliding window and the word level are even lower levels and are used to calculate the strengths of word associations between words. These low-level calculations are also used in higher levels, but are combined to achieve all different association directions.

B. Named Entities

Named Entity Recognition (NER) can provide meaningful information based on text analysis and applied text mining methods [16]. Entities which have been recognized in textual content have associated knowledge and therefore, should be semantically referenced to other information in the knowledge base. NER uses Part-of-Speech (POS) to identify proper nouns. Based on the sentence sequences and training of the deciding model, the probability of each POS tag can be expressed. Especially for proper nouns, the knowledge base can be queried for existing knowledge to add possible newly extracted knowledge facts to the knowledge base in the semantically correct way. In dimensional structured knowledge base, the entities can be used as a selection tool for associative and contextual information.

IV. Contextual search of associated information

Searching for information is often a very time consuming task. Unstructured document-based repositories often only provide a keyword-based search on the content which may result in a large number of search results. For employees, it is even more difficult to search inside the first filtered results for their requested information. Furthermore, complex structures and cryptic file names of documents require inspecting a large amount of the textual resources. Moreover, the growing number of documents and produced data leads to a more and more challenging task to find the correct and required information. In organizations, these problems can quickly result into being real cost drivers and the existing knowledge potentials remain unused.

To be efficient, employees need a much better search methodology than specifying one or more keywords. An intelligent searching approach is the provision of contextual information within the search engine by utilizing word associations for suggestion of (additional) search terms.

In many available search engines, the next term suggestions are done based on frequently used search expressions from (other or similar) users. The contextual search of associated information follows a different approach. By consideration of the word association strengths within the content, closer related keywords can be automatically retrieved and suggested for further filtering of search results. After the first keyword is specified in the search engine and the first results are provided, all associated terms are identified by the calculation of CIMAWA word association strength. The highest strength of word association indicates the closest (semantic) relation and therefore additional words are suggested as a possible additional term. In this way, the search results from the search engine are further filtered by directly taking additional associated knowledge into account. Finally, the information is extracted from the knowledge base and presented in the knowledge map, a form that is visually understandable by a broad spectrum of users.

A. Discovery of related content based on the CIMAWA

CIMAWA models the mental association of the human mind. The hybrid association measure can be applied in different areas and has been utilized for many purposes. In the following, a practical implementation of the CIMAWA based associative search is described in detail.

The basis for the calculation of the word associations are the textual resources within the dimensional structured knowledge base. The text from stored documents and reports must be extracted into the knowledge base. They are applicable pre-processed in a first step, as required by the application. The

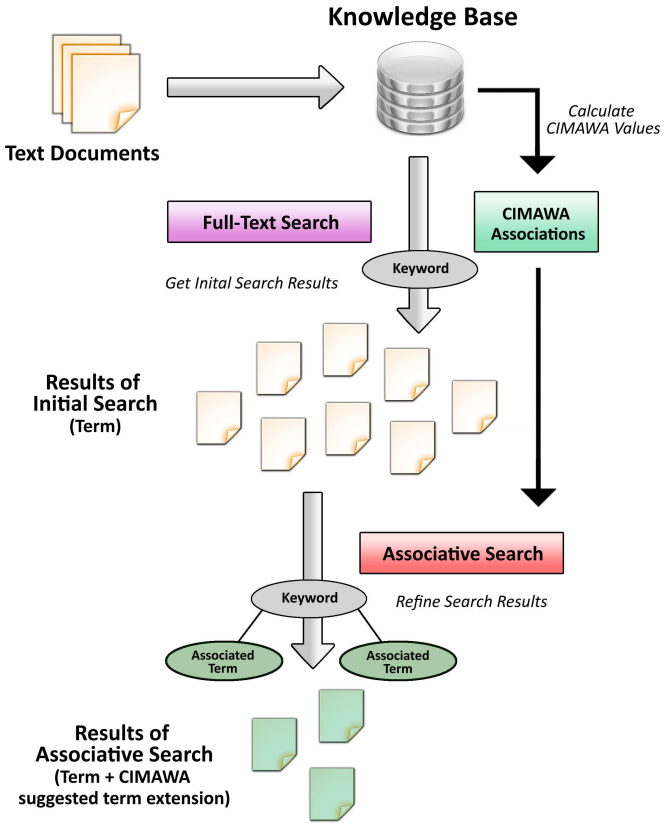

Fig. 1. Visualization of associative search with associative term extensions based on the CIMAWA

pre-processing is done in the background and doesn't require any interaction with the user of the searching system.

The first interaction with the system is the input of the request in the form of one or more search terms. This request, or more precisely the term entered, is further processed by the system in two ways.

1) Full-Text Search: The entire textual content of the knowledge base is searched as part of a full-text search for all entered terms. The obtained results in the form of the text passages which contain the terms are offered for the user as search results. These results provide the direct matched results from the knowledge base according to the entered term. Depending on the entered term and depending on the size of the knowledge base, a large number of direct search results is found in the full-text search and the user is forced to carry out a time-consuming manual search in order to reach the relevant textual information.

In order to filter the number of results, the associative search provides different terms as search expression extension to add further keywords suggested associated, contextual information.

2) Associative Term Extension: The textual knowledge base is analyzed and the values of the word association strength between all the terms are calculated. In this way, semantically-related terms are recognized and offered as term extensions to the user in order to refine the search. Each of the offered terms can be selected by the user to further specify the user's contextual requirements. The user selects, depending on the current task, contextual information from a displayed list. Since the association calculation is performed based on the company's internal knowledge base, which simultaneously represents the text collection, it is ensured that each term occurs simultaneously in one or more text passages of the knowledge base. This means a real added value for the user, because behind every offered term extension is available information in the form of further text passages. Consequently, the selection reduces the amount of the direct search results of full-text search to an appropriate subset of text passages with the selected term extension. A visualization of the associative search concept with associative term extensions based on the CIMAWA is illustrated in Figure 1.

For example, if initially searched for a specific product in the product portfolio, the full-text search will find numerous search results from different sources. With the aid of semantically-related terms, such as a special product name or even a customer, the number of search hits is greatly decreased and the relevance of these results is highly increased. By following this approach, the advantage is that the word associations are calculated based on the company's internal knowledge base and occur in the direct context of the search term. Results of prototype implementation of the associative search have shown that the suggested terms also include terms which are unknown outside of the organization. Expressions which are only used internally in the knowledge base, like different product names or descriptions which have evolved over time within the knowledge base are also included.

B. Associative proximity measure

The CIMAWA calculations, previously mentioned in the associative search will result in a multitude of word association strength values. Based on these calculated values, CIMAWA can be applied as a measurement for graphical representation of search results. The calculated values can be interpreted as distance vectors between different terms in the knowledge base. In this way, more related content can be visualized closer to the initial search term, which is displayed as center of the graph. For the calculation of the distances between different terms the following equation, presented in [7], can be used to calculate the word association strengths and provides the numeric values which are used to describe the distances.

$$CIMAWA^{\zeta}_{ws}(x(y)) = \frac{Cooc_{ws}(x,y)}{(frequency(y))^{\alpha}} + \zeta \cdot \frac{Cooc_{ws}(x,y)}{(frequency(x))^{\alpha}} \quad (1)$$

In Equation (1), the variable x is used for keywords, especially named entities, within the knowledge map which have been extracted from the textual resource. For variable y, different terms can be used and the CIMAWA calculation is iteratively done for each of the other terms within the knowledge base. By utilizing the co-occurrences $Cooc_{ws}(x, y)$ with text window size ws (normally a window of five words is

taken), all word associations strengths between x and different terms for y are provided. In this way, an associative distance from the named entity, which is word x, can be specified for all other terms in the knowledge base. By applying the calculations on multiple entities, relations between entities through communally associated terms are identified and can be used for the distance description and layout structuring of the knowledge map. Since the calculations need to be re-done from time to time (after new knowledge has been inserted into the knowledge base which possibly brings new insights) the distance values are changing over the time. This is due to the fact, that the association of content is distinguished by the content itself. The re-calculations don't necessarily need to be re-done after every textual resource insertion for the whole knowledge map. At a certain point, after the knowledge base contains enough textual information, the CIMAWA values will remain almost stable. Moreover, only the affected entities and terms which occur in the new textual resource need to be re-calculated. The next section described how the re-calculation over the time can be used to provide a fully dynamic, self-adjusting knowledge map.

V. Knowledge Discovery of Contextual Content

Knowledge discovery is a challenging task if a knowledge base doesn't offer an interface with the possibilities to easily browse and search for specific information. In organizations, the knowledge base is often considered as database, distributed databases or cloud storage which support different applications by provision of necessary data. However, if the knowledge base of a company is considered only as a static repository of documents or knowledge assets, the discovery of knowledge is a very time-consuming task.

The dimensional structured knowledge base design offers possibilities for knowledge discovery purposes. The main advantages and characteristics of multidimensional knowledge bases are scalability, flexibility, dimensionality and relevance [6]. This means, that entities like persons, brands, locations can be selected from the knowledge base based on dimensional relationships and used for analysis.

As an example, different persons (named entities) can be selected from the knowledge base and all related facts about the persons are provided. In this scenario, an associative knowledge map could visualize all facts based on the word association strength. Because of the dimensional structuring of the knowledge base, this knowledge map could be further re-arranged based on additional dimensional selection. In the aforementioned example, the associative knowledge map would only visualize the facts about persons regarding specific topics (if topics are selected as further dimensional information).

In the following, the dynamic knowledge map is described shortly. Afterwards, a use case for the dynamic knowledge map in virtual reality applications is mentioned.

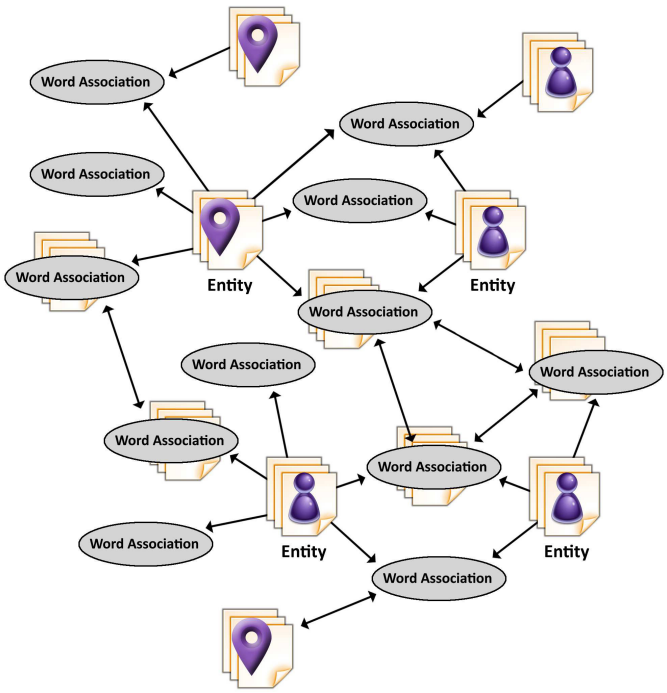

Fig. 2. Conceptual overview of the dynamic knowledge map. Different entities (e.g. persons, places) are arranged by distances derived from CIMAWA word association strength.

A. Dynamic Knowledge Maps

Knowledge maps are able to provide a visual summarization of information and the relations between each piece of information. Based on graphical arrangement, different nodes, symbols or graphical illustrations related to the knowledge assets or pieces of information can be used in the knowledge map. Entities which have been identified by NER are used in the dynamic knowledge map to improve the graphical layout and visual understandability. Based on the type of named entity (e.g. product, person, location), different (standard) icons can be used in the knowledge map to differentiate between them. Another way is the automatic retrieval of related images from the web in order to visualize the entity information inside the nodes within the knowledge map.

The proximity calculations based on the CIMAWA measure which have been explained in the previous section can be used for the arrangement of the content. Associative information is visualized next to related knowledge items whereas unrelated information is hidden or visualized with higher distance. As previously explained, higher distance implies a none existing or lower word association strength based on CIMAWA.

Further functionality like zooming in and out of the map is one of the main objectives of a dynamic knowledge map. For this characteristic, a further break down of summarized information is necessary. For textual resources, it means that documents are splitted into their sections, pages, paragraphs or even sentences. The associations between the contained keywords within the text fragments and desired detail level

are able to provide the information for graphical arrangement based on word association strength. In this way, the dynamic knowledge map is able to restructure itself and adapt to the chosen information depth. A conceptual overview of the dynamic knowledge map is depicted in Figure 2.

B. Virtual and Augmented Reality Applications

Virtual Reality (VR) and Augmented Reality (AR) will revolutionize the creation, presentation, testing and experience of products and services in future. Besides the entertainment aspects of VR and AR scenarios, different realities will also influence the way we produce, capture and store information in daily work life. In this regard, an information and knowledge provider is necessary to access contextual information in different applications and situations. The dynamic knowledge map is one of the tools that offer a graphical structure which could be transferred into different VR/AR supported applications. In this way, the graphical structure of the dynamic knowledge map can be directly explored in those environments. To this aim, the knowledge map has to be continuously organized by the current situation and context of the user. Moreover, the map also needs to provide an access to the desired information in real-time.

In organizations, VR/AR applications, user interfaces within and the way information is presented need to be adapted according to the new forms of job and organizational environments. The dynamic knowledge map is one possible solution for provision of additional information and can be virtually transferred into an explorable environment which provides contextual information based on the current user interaction.

VI. Conclusion

In this paper we have shown how knowledge maps can be utilized as a visualization instrument for dimensional structured associative knowledge. By bringing textual information in a knowledge base, documents can be analyzed on different levels. Each of the levels has different other associated knowledge and therefore a different possible context. The multidimensional knowledge base has the capability to support the different information levels in terms of knowledge scalability. Especially the textual content can be summarized and further splitted into details based on dimensional structures.

The associative search is a useful method to select relevant contextual information from the knowledge base. It can be implemented in different use case applications. In future work we are planning to extend the concept and further implement it as a knowledge base search methodology in various use cases.

Moreover, an approach for dynamic knowledge maps based on word association strengths has been presented in this paper. By the utilization of word association strengths as a proximity measurement for layout creation, the possibilities for the visualization of knowledge map have been further described. The automatic layout planning and optimization of knowledge maps provided word association strengths will be also a further work in this area.

References

[1] A. Tiwana. The knowledge management toolkit: practical techniques for building a knowledge management system. Prentice Hall PTR, 2000.

[2] A. Balaid, M. Z. A. Rozan, S. N. Hikmi, and J. Memon. Knowledge maps: A systematic literature review and directions for future research. International Journal of Information Management, 36(3), 451-475, 2016.

[3] A. M. O'donnell, D. F. Dansereau, and R. H. Hall. Knowledge maps as scaffolds for cognitive processing. Educational Psychology Review, 14(1), 71-86, 2002.

[4] M. J. Eppler. Making knowledge visible through intranet knowledge maps: concepts, elements, cases. In System Sciences, 2001. Proceedings of the 34th Annual Hawaii International Conference on (pp. 9-pp), IEEE, 2002

[5] M. J. Eppler. A process-based classification of knowledge maps and application examples. Knowledge and Process Management, 15(1), 59-71, 2008.

[6] J. Zenkert, and M. Fathi. Multidimensional Knowledge Representation of Text Analytics Results in Knowledge Bases. 2016 IEEE International Conference on Electro/Information Technology (EIT), North Dakota, USA, 2016, In press.

[7] P. Uhr, A. Klahold, and M. Fathi. Imitation of the human ability of word association. International Journal of Soft Computing and Software Engineering (JSCSE), 3(3), 2013.

[8] F. R. Lin, and C. M. Hsueh. Knowledge map creation and maintenance for virtual communities of practice. Information Processing & Management, 42(2), 551-568, 2006.

[9] T. H. Ong, H. Chen, W. K. Sung, and B. Zhu. Newsmap: a knowledge map for online news. Decision Support Systems, 39(4), 583-597, 2005.

[10] J. Watthananon, and A. Mingkhwan. Optimizing knowledge management using knowledge map. Procedia Engineering, 32, 1169-1177, 2012.

[11] J. H. Woo, M. J. Clayton, R. E. Johnson, B. E. Flores, and C. Ellis. Dynamic Knowledge Map: reusing experts' tacit knowledge in the AEC industry. Automation in construction, 13(2), 203-207, 2004.

[12] H. R. Suresh, and C. O. Egbu. Knowledge mapping: concepts and benefits for a sustainable urban environment. In 20th Annual Conference Association of Researchers in Construction Management (ARCOM) (pp. 1-3), 2004.

[13] W. Vestal. Knowledge Mapping: The Essentials for Success. APQC, p.75, 2005.

[14] Y. Dang, Y. Zhang, H. Chen, and C. A. Larson. Knowledge Mapping for Bioterrorism-Related Literature. In Infectious Disease Informatics and Biosurveillance (pp. 311-338), Springer US, 2011.

[15] A. Klahold, P. Uhr, F. Ansari, and M. Fathi. Using Word Association to Detect Multitopic Structures in Text Documents, IEEE Intelligent Systems, vol. 29, no. 5, pp. 40-46, Sept.-Oct. 2014.

[16] D. Nadeau, and S. Sekine. A survey of named entity recognition and classification. Lingvisticae Investigationes, 30(1), 3-26, 2007.