Metadata of the chapter that will be visualized in SpringerLink

Book Title	Data Management Technologies and Applications	
Series Title		
Chapter Title	An Automatic Construction of Concept Maps Based on Statistical Text Mining	
Copyright Year	2016	
Copyright HolderName	Springer International Publishing Switzerland	
Corresponding Author	Family Name	**Nugumanova**
	Particle	
	Given Name	**Aliya**
	Prefix	
	Suffix	
	Division	
	Organization	D. Serikbayev East Kazakhstan State Technical University
	Address	Ust-Kamenogorsk, Kazakhstan
	Email	yalisha@yandex.kz
Author	Family Name	**Mansurova**
	Particle	
	Given Name	**Madina**
	Prefix	
	Suffix	
	Division	
	Organization	Al-Farabi Kazakh National University
	Address	Almaty, Kazakhstan
	Email	mansurova01@mail.ru
Corresponding Author	Family Name	**Alimzhanov**
	Particle	
	Given Name	**Ermek**
	Prefix	
	Suffix	
	Division	
	Organization	Al-Farabi Kazakh National University
	Address	Almaty, Kazakhstan
	Email	aermek81@gmail.com
Author	Family Name	**Zyryanov**
	Particle	
	Given Name	**Dmitry**
	Prefix	
	Suffix	
	Division	
	Organization	D. Serikbayev East Kazakhstan State Technical University
	Address	Ust-Kamenogorsk, Kazakhstan

	Email	dzyryanov@ektu.kz
Author	Family Name	**Apayev**
	Particle	
	Given Name	**Kurmash**
	Prefix	
	Suffix	
	Division	
	Organization	D. Serikbayev East Kazakhstan State Technical University
	Address	Ust-Kamenogorsk, Kazakhstan
	Email	kapaev@ektu.kz

Abstract	In this paper, we explore the task of automatic construction of concept maps for various knowledge domains. We propose a simple 3-steps algorithm for extraction of key elements of a concept map (nodes and links) from a given collection of domain documents. Our algorithm manipulates a statistical term-document matrix describing how frequently terms occur in documents of the collection. At the first step we decompose this matrix into scores (terms-by-factors) and loadings (factors-by-documents) matrixes using non-negative matrix factorization, wherein each factor represents one topic of the collection. Since the scores matrix specifies the relative contribution of each term to the factors, we can select the most contributing terms and use them as concept map nodes. At the second step we associate selected key terms with the corresponding row-vectors of the term-document matrix and calculate pairwise cosine distances between them. Since the close distances determine the pairs of strongly related key terms, we can select the strongest relations as concept map links. Finally, we construct the resulting concept map as a graph with selected nodes and links. The benefits of our statistical algorithm are its simplicity, efficiency and applicability to any domain, any language and any document collection.

Keywords (separated by '-')	Concept map - Text mining - Co-occurrence analysis - Non-negative matrix factorization

An Automatic Construction of Concept Maps Based on Statistical Text Mining

Aliya Nugumanova[1(✉)], Madina Mansurova[2], Ermek Alimzhanov[2(✉)], Dmitry Zyryanov[1], and Kurmash Apayev[1]

[1] D. Serikbayev East Kazakhstan State Technical University, Ust-Kamenogorsk, Kazakhstan
yalisha@yandex.kz, {dzyryanov,kapaev}@ektu.kz
[2] Al-Farabi Kazakh National University, Almaty, Kazakhstan
mansurova01@mail.ru, aermek81@gmail.com

Abstract. In this paper, we explore the task of automatic construction of concept maps for various knowledge domains. We propose a simple 3-steps algorithm for extraction of key elements of a concept map (nodes and links) from a given collection of domain documents. Our algorithm manipulates a statistical term-document matrix describing how frequently terms occur in documents of the collection. At the first step we decompose this matrix into scores (terms-by-factors) and loadings (factors-by-documents) matrixes using non-negative matrix factorization, wherein each factor represents one topic of the collection. Since the scores matrix specifies the relative contribution of each term to the factors, we can select the most contributing terms and use them as concept map nodes. At the second step we associate selected key terms with the corresponding row-vectors of the term-document matrix and calculate pairwise cosine distances between them. Since the close distances determine the pairs of strongly related key terms, we can select the strongest relations as concept map links. Finally, we construct the resulting concept map as a graph with selected nodes and links. The benefits of our statistical algorithm are its simplicity, efficiency and applicability to any domain, any language and any document collection.

AQ1

Keywords: Concept map · Text mining · Co-occurrence analysis · Non-negative matrix factorization

1 Introduction

Concept maps are graphical tools for representing knowledge structures of various domains. The main elements of concept maps are:

- Nodes (key concepts of the domain, put in circles or boxes);
- Links (key relations between concepts, represented as lines);
- Labels (words or phrases describing the meaning of relations).

The main purpose of concept maps is to contribute to a deeper understanding of domain knowledge on the conceptual level. The work [1] reports the results of experimental investigations which verify the practical value and efficiency of concept

M. Helfert et al. (Eds.): DATA 2015, CCIS 584, pp. 1–10, 2016.
DOI: 10.1007/978-3-319-30162-4_3

maps as a tool and as a strategy of teaching. Unfortunately, the complexity of a manual construction of concept maps greatly reduces the advantages of their using in the educational process. It is a very common case when teachers preparing course material are forced to use simple and limited types of concept maps or not use them at all because their comprehensive construction and drawing takes a lot of time.

Owing to the mentioned arguments, of great importance is the task of automatic or semi-automatic construction of concept maps on the basis of extraction their elements from collections of textual materials. Thanks to example of the authors of work [2] this task got the name Concept Map Mining (CMM) similar to Data Mining and Text Mining. In general case the process of CMM consists of three subtasks: extraction of concepts, extraction of links and summarization (see Fig. 1) [3].

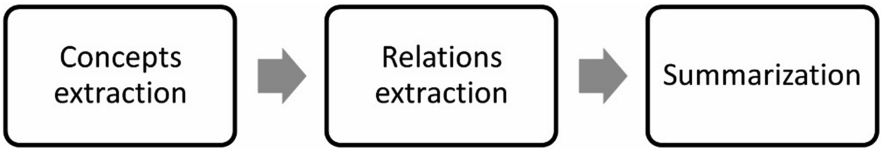


Fig. 1. The subtasks of concept map mining process.

The aim of this paper is to demonstrate usefulness and efficiency of statistical Text Mining methods for automatic construction of concept maps based on the domain collections of texts. The most important advantage of statistical methods is that they can be directly applied to any domain and any language, i.e. they are invariant in regard to the features of the given collection of domain documents. Statistical methods considered in this paper are based on the analysis of co-occurrence of terms in domain documents. We use a simple 3-steps algorithm which deals with a term-document co-occurrence matrix describing the number of occurrences of each word in each document of the collection.

At the first step we decompose the co-occurrence matrix into scores (terms-by-factors) and loadings (factors-by-documents) matrixes using non-negative matrix factorization, wherein each factor represents one topic of the collection. Since the scores matrix specifies the relative contribution of each term to the factors, we can select the terms with maximum contributions and use them as concept map nodes. At the second step we associate selected key terms (also known as concepts or nodes) with the corresponding row-vectors of the term-document matrix and calculate pairwise cosine distances between them. Cosine distance or similarity is a measure of similarity between two vectors that measures the cosine of the angle between them. We use this measure in positive space, where the outcome is bounded in [0,1]. So the maximum value of cosine similarity is equal to 1 (it corresponds to the angle 0). Since the close distances determine the pairs of strongly related key terms, we can select the strongest relations with similarity values more than 0.5 as concept map links. Finally, we construct the resulting concept map as a graph with selected nodes and links.

We plan to integrate automatically created concept maps into e-learning environment as a special tool supporting student's active and deep learning of the subject.

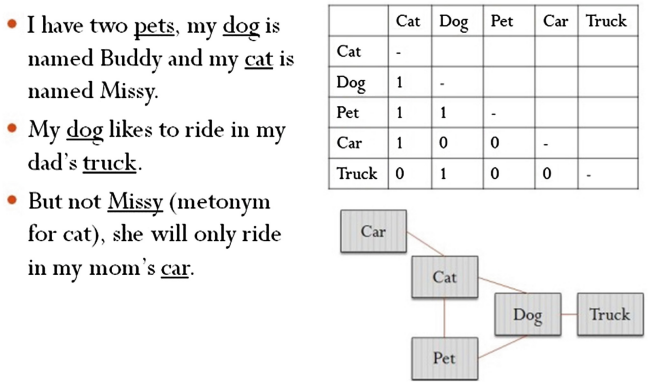
In [4] we represent the conception of our e-learning environment, and in this paper we investigate one of its meaningful elements.

The remaining part of the paper has the following structure. The second section presents a brief review of works related to the considered problem of automatic construction of concept maps. The approach proposed in the paper is described in detail in the third section. The results of experimental testing of the proposed approach are given in the fourth section. The fifth section contains brief conclusions on the work done and presents a plan of further investigations.

2 Related Works

The resent decade is characterized by the growth of interest to investigations devoted to automatic extraction of concept maps from collections of text materials. Among these studies, of high rank are the works based on the use of statistical techniques of processing a natural language. As is mentioned in [5], the methods focused on statistical processing of texts are simple, efficient and well portable; however, they possess a decreased accuracy as they do not consider latent semantics in the text.

The mentioned simplicity and efficiency of statistical approaches are illustrated well in [6]. The authors construct a term-term matrix based on a short list of key words selected manually for the given domain. They fill in the matrix on the basis of terms co-occurrences in sentences. If two elements occur in one sentence, the matrix element is equated to 1, otherwise – to 0. Then they display this matrix in the concept map, as shown in Fig. 2. Obviously, this approach is good for chamber teaching courses consisting of materials limited in volume, but for weighty courses it is very inefficient. The authors applied their methodology for constructing concept maps based on students' text summaries. Obtained concept maps were used by instructor to analyze how students learned the training material. In particular, the purpose of the analysis was selection of correct, incorrect or missing propositions in the students' summaries.



	Cat	Dog	Pet	Car	Truck
Cat	-				
Dog	1	-			
Pet	1	1	-		
Car	1	0	0	-	
Truck	0	1	0	0	-

Fig. 2. Mapping the term-by-term matrix to the concept map (by work [6]).

The authors of [7] extract concepts from scientific articles using the principal component analysis. They use some papers in scientific journals and conference proceedings, dedicated to the field of e-learning, as data sources for the construction of concept maps. According to them, constructed concept maps can be useful for researchers who are beginners to the field of e-learning, for teachers to develop adaptive learning materials, and for students to understand the whole picture of e-learning domain. The authors introduce the notion "relation strength" with the help of which they describe pairwise relations between extracted concepts. Relation strength is calculated on the basis of distance between two concepts in the text and on the basis their co-occurrence in one sentence. Authors link pairs of concepts which have "relation strength" more than 0.4. Like the authors of [6], the authors of this work do not label found links (do not sign them).

Generally speaking, labeling of relations extracted from the texts is a very complex problem that requires performing semantic analysis of texts. That is why many researchers note the limitedness of statistical approaches and try to combine statistical and linguistic tools by using knowledge bases suitable for semantic analysis. For example, the authors of [8] use thesaurus WordNet for part-of-speech analysis of texts. Due to determination of parts of speech in sentences they extract a predicate (the main verb) from each sentence and form for each predicate a triplet "subject-predicate-object". The subject and object are interpreted as concepts and the predicate as a relation between them (see Fig. 3). The authors of the paper are interested in building a concept map concerning biological kingdoms.

Fig. 3. "Subject-predicate-object" triplet used in [8].

The authors of [9] analyze the structure of sentences by constructing trees of dependences. They divide each sentence into a group of members dependent on the noun and a group of members dependent on the verb. They display verbs in the links and the nouns in concepts, as shown in Fig. 4. The final goal of the authors is to develop intelligent user interfaces to help understanding of complex project documents

and contextualization of project tasks. The paper [10] describes an approach based on the use of thesaurus WordNet, too. The authors of this work use the lexical power of WordNet to provide the construction of an interactive concept maps by students. Using WordNet, the authors perform processing of different student responses revealing the meaning of the concepts with the help of synonyms hyponyms, meronyms, and homonyms existing in the lexical base of WordNet.

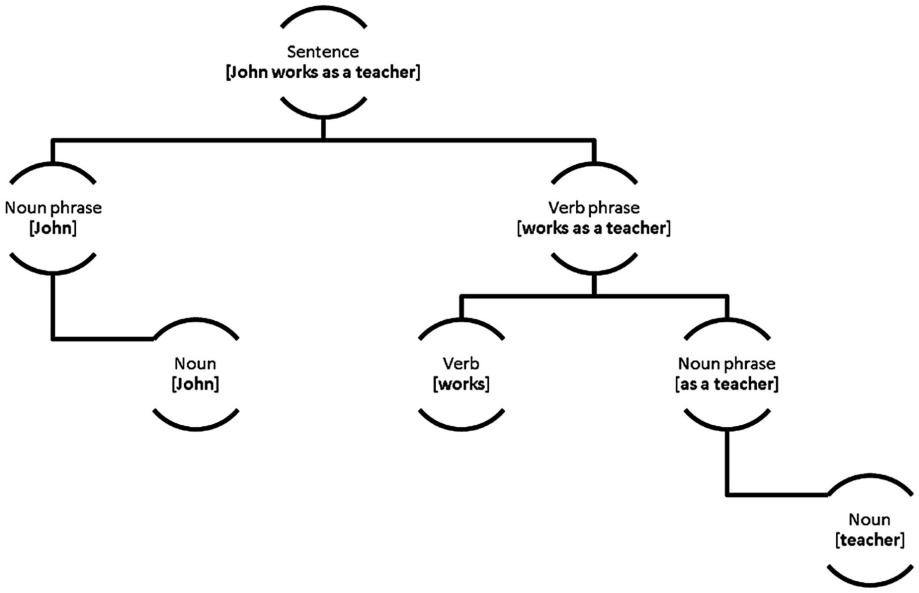


Fig. 4. Dependency tree for semantic analysis used in [9].

Like the authors of [6], the authors of [11] search for "noun-verb-noun" structures in sentences. They use verbs as designations of links and display nouns in concepts. The authors of [12] use not only verbs but also prepositional groups of the English language which designate possessiveness (of), direction (to), means (by), etc. for designation of links. The authors of [13] propose a novel approach based on combined techniques of automatic generation of exhaustive syntactic rules, restricted-context part-of-speech tagging and vector space intersection. They start from a basic set of simple syntactic rules (Noun-Verb-Noun, Verb-Noun-Verb) and expand the concept of noun (N) to include other syntagmatic components.

All the enumerated works demonstrate quite good results for extraction of concepts and relations. The problem only occurs when marking relations, i.e. when assigning semantics to relations. Interpretation of verbs and prepositional groups as relations is one of the ways to solve this problem which requires the use of linguistic tools and dictionaries.

3 Proposed Approach

3.1 Concepts Extraction

The first step of our approach is extraction of domain key terms which can be used as concepts – basic elements of a concept map. We start this step with preprocessing of a given textual collection, i.e. division texts into words, lemmatization (reduction of words to normal forms) and removal of stop-words. As result of such preprocessing we obtain a list of unique words (terms) of the collection. After that we construct a term-document matrix the rows of which correspond to terms, columns – to documents and elements – to frequencies of using terms in documents.

We decompose the obtained co-occurrence matrix into scores (terms-by-factors) and loadings (factors-by-documents) matrixes using non-negative matrix factorization, wherein each factor represents one topic of the collection (see Fig. 5). Non-negative matrix factorization is a very fruitful technique used for dimensionality reduction [14]. It produces data projections in the new factor space wherein each factor is represented as a vector of relative contributions of terms. We can sort terms of each factor by their contributions in descending order, select first p elements and generate a list of dominant terms. Conjunction of all lists gives a final list of dominant domain terms (concepts).

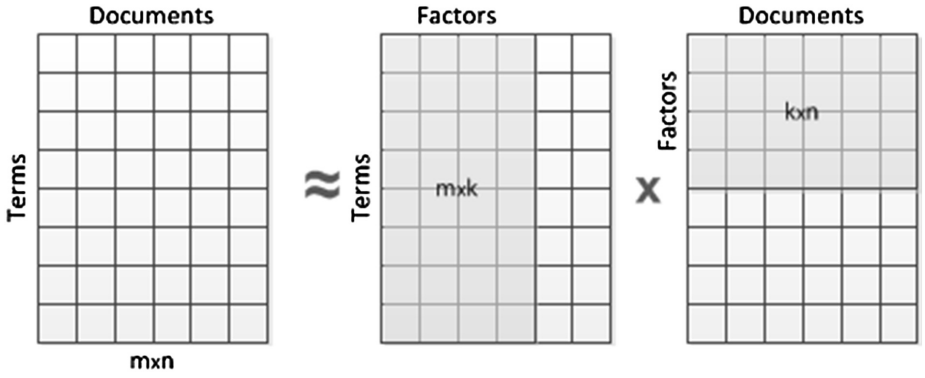


Fig. 5. Non-negative matrix factorization.

3.2 Relations Extraction

The obtained term-document matrix contains information concerning links between all terms and all documents. To extract relations between concepts, we should concentrate on links between selected key terms. So we should exclude from our term-document matrix rows which do not correspond to key terms selected on the previous step. Thereby we should reduce the dimension of our matrix. Then we should transform this reduced term-document matrix to a term-term matrix. For this, we should find pairwise distances between rows of the term-document matrix. The distance can be calculated using the cosine measure:

$$c = \cos(\overline{x}, \overline{y}) = \frac{\overline{x} \cdot \overline{y}}{|x| \cdot |y|}$$

where c is a distance value; x, y are any two rows in the reduced term-document matrix corresponding to the pair of concepts. The obtained values are measured by figures in the range from 0 to 1. The higher the similarity between vectors-terms, the less is the angle, the higher is the cosine of the angle (cosine measure). Consequently, maximum similarity is equal to 1, and minimum one is equal to 0.

The obtained term-term matrix measures distances between terms based on their co-occurrence in documents (as coordinates of vectors-terms are frequencies of their use in documents). It means that the sparser the initial term-document matrix, the worse is the quality of the term-term distances matrix. Therefore, it is expedient to save the initial matrix from information noise and rarefaction with the help of the latent semantic analysis [15]. The presence of noise is conditioned by the fact that, apart from the domain knowledge, initial documents contain "general places" which, nevertheless, contribute to the statistics of distribution.

We use the method of latent semantic analysis for clearing up the matrix from information noise. The essence of the method is based on approximation of the initial sparse and noised matrix by a matrix of lesser rank with the help of singular decomposition. Singular decomposition of matrix A with dimension $M \times N$, $M > N$ is its decomposition in the form of product of three matrices – an orthogonal matrix U with dimension $M \times M$, diagonal matrix S with dimension $M \times N$ and a transposed orthogonal matrix V with dimension $N \times N$:

$$A = USV^T \tag{1}$$

Such decomposition has the following remarkable feature. Let matrix A be given for which singular decomposition $A = USV^T$ is known and which is needed to be approximated by matrix A_k with the pre-determined rank k. If in matrix S only k greatest singular values are left and the rest are substituted by nulls, and in matrices U and V^T only k columns and k lines are left, then decomposition

$$A_k = U_k S_k V_k^T \tag{2}$$

will give the best approximation of the initial matrix A by matrix of rank k. Thus, the initial matrix A with the dimension $M \times N$ is substituted with matrices of lesser sizes $M \times k$ and $k \times N$ and a diagonal matrix of k elements. In case when k is significantly less than M and N, we have a significant compression of information. However, part of information is lost and only the most important (dominant) part is saved. The loss of information takes place because of neglecting small singular values, i.e. the more singular values are discarded the higher the loss. Thus, the initial matrix gets rid of information noise introduced by random elements.

3.3 Summarization

The extracted concepts and relations must be plotted on a concept map. Let us repeat that as concepts we use terms which contribution to collection factors is higher than a certain threshold value determined experimentally. Varying this value, we can reduce or increase the list of concepts. In the same way, we can vary the number of extracted relations. Among all pairwise distances in the term-term matrix we select the values higher than a certain threshold. Therefore, we select only edges (links) which connect only the concepts the proximity between which is higher than the indicated threshold.

4 Experiments

To carry out experiments, we chose the subject domain "Ontology engineering". The documents representing chapters from the textbook [16] formed a teaching collection. Tokenization and lemmatization from the collection resulted in a thesaurus of unique terms. The use of non-negative matrix factorization allowed selecting 500 key concepts of the subject domain. Table 1 presents the first 12 concepts.

Table 1. Key extracted concepts.

No	Concept
1	Semantic
2	Web
3	Property
4	Manner
5	Model
6	Class
7	Major
8	Side
9	Word
10	Query
11	Rdftype
12	Relationship

Then the constructed term-document matrix was approximated by a matrix of the rank 100 with the help of singular decomposition. On the basis of the obtained matrix, pairwise distances between terms-lines were calculated using cosine measure. Thus, the transfer from a term-document matrix to a term-term matrix was carried out. Table 2 presents, as an example, some pairs of terms with different indexes of proximity. Only the links the proximity values of which exceeded 0.5 were left as relations significant for construction of a concept map.

Having obtained all concepts and links, we constructed a graph of the concept map. The concepts were taken as nodes of the graph and relations between concepts were taken as edges. As the general structure of the map is too large for analysis, we present a fragment of this map in Fig. 6.

AQ2

Table 2. The samples of various extracted relations.

No	First concept	Second concept
1	OWL	Class
2	OWL	Modeling
3	OWL	Member
4	Property	Class
5	Result	Pattern
6	Term	Relationship

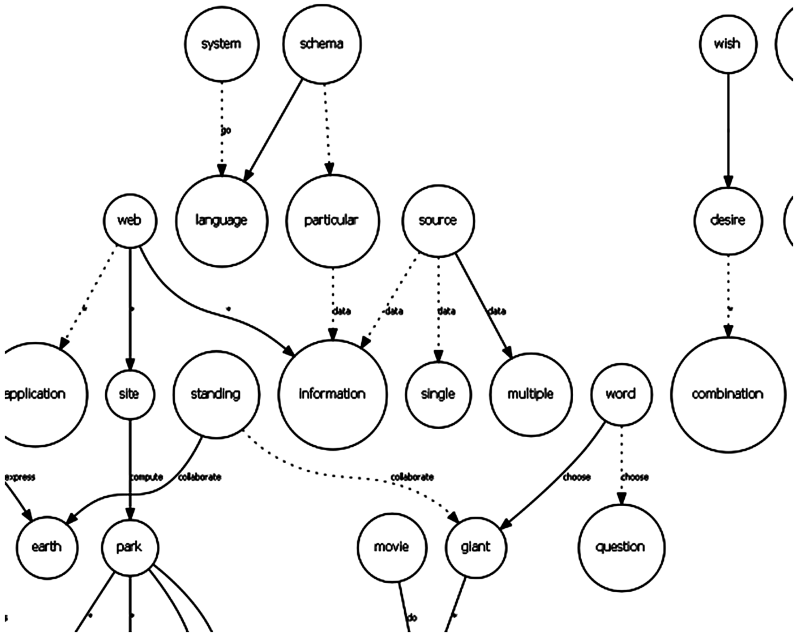


Fig. 6. Fragment of the concept map.

5 Conclusion

We are introduced another method of constructing concept maps and experimental results have been positively evaluated by two independent experts in the domain. Further studies will be related with the processing of large concept maps, their visualization and intelligent processing methods.

This work is part of a project carried out in the Al-Farabi Kazakh National University, the goal of which is to develop efficient algorithms and models of semi-structured data processing, on the basis of modern technologies in the field of the Semantic Web using the latest high-performance computing achievements to obtain new information and knowledge from unstructured sources, large amounts of scientific data and texts.

References

1. Sherman, R.: Abstraction in concept map and coupled outline knowledge representations. J. Interact. Learn. Res. **14**, 31–49 (2003)
2. Villalon, J., Calvo, R.: Concept map mining: a definition and a framework for its evaluation. In: Proceedings of the International Conference on Web Intelligence and Intelligent Agent Technology, vol. 3, Los Alamitos, USA, pp. 357−360 (2008)
3. Villalon J., Calvo R., Montenegro R.: Analysis of a gold standard for Concept Map Mining – how humans summarize text using concept maps. In: Proceedings of the Fourth International Conference on Concept Mapping, pp. 14−22 (2010)
4. Akhmed-Zaki D., Mansurova M., Pyrkova A.: Development of courses directed on formation of competences demanded on the market of IT technologies. In: Proceedings of the 2014 Zone 1 Conference of the American Society for Engineering Education, pp. 1−4 (2014)
5. Zubrinic, K., Kalpic, D., Milicevic, M.: The automatic creation of concept maps from documents written using morphologically rich languages. Expert Syst. Appl. **39**(16), 12709–12718 (2012)
6. Clariana, R.B., Koul, R.: A computer-based approach for translating text into concept map-like representations. In: Proceedings of the First International Conference on Concept Mapping, Pamplona, Spain, pp. 131–134 (2004)
7. Chen, N.S., Kinshuk Wei, C.W., Chen, H.J.: Mining e-learning domain concept map from academic articles. Comput. Educ. **50**(3), 1009–1021 (2008)
8. Oliveira, A., Pereira, F.C., Cardoso, A.: Automatic reading and learning from text. In: Paper Presented at the International Symposium on Artificial Intelligence Kolhapur, India (2001)
9. Valerio, A., Leake, D.B.: Associating documents to concept maps in context. In: Paper Presented at the Third International Conference on Concept Mapping, Finland (2008)
10. Alves, Z.O., Pereira, F.C., Cardoso, A.: Automatic reading and learning from text. In: Proceedings of the International Symposium on Artificial Intelligence (ISAI 2001), pp. 302–310 (2001)
11. Rajaraman, K., Tan, A.H.: Knowledge discovery from texts: a concept frame graph approach. In: Proceedings of the 11th International Conference on Information and Knowledge Management, pp. 669–671 (2002)
12. Valerio, A., Leake, D.B., Cañas, A.J. Using automatically generated concept maps for document understanding: a human subjects experiment. In: Proceedings of the 15 International Conference on Concept Mapping, pp. 438−445 (2012)
13. Reis, J.C., Gaia, A.S.C., Viegas Jr, R.: Concept maps construction based on exhaustive rules and vector space intersection. IJCSNS **14**(7), 26 (2014)
14. Costa, G., Ortale, R., A latent semantic approach to xml clustering by content and structure based on non-negative matrix factorization. In: 2013 12th International Conference on Machine Learning and Applications (ICMLA) IEEE 2013, vol. 1, pp. 179−184 (2013)
15. Evangelopoulos, N.E.: Latent semantic analysis. Wiley Interdisc. Rev.: Cognitive Sci. **4**(6), 683–692 (2013)
16. Allemang, D., Hendler, J.: Semantic Web for the Working Ontologist, 2nd edn. Elsevier Inc., Philadelphia (2011)

Author Query Form

Book ID :	**418335_1_En**
Chapter No.:	**3**

Springer
the language of science

Please ensure you fill out your response to the queries raised below and return this form along with your corrections

Dear Author
During the process of typesetting your chapter, the following queries have arisen. Please check your typeset proof carefully against the queries listed below and mark the necessary changes either directly on the proof/online grid or in the 'Author's response' area provided below

Query Refs.	Details Required	Author's Response
AQ1	Please confirm if the corresponding authors are correctly identified. Amend if necessary.	
AQ2	The citation of Fig. 8 has been changed to Fig. 6 in the citation. Please check and conform.	

MARKED PROOF

Please correct and return this set

Please use the proof correction marks shown below for all alterations and corrections. If you wish to return your proof by fax you should ensure that all amendments are written clearly in dark ink and are made well within the page margins.

Instruction to printer	*Textual mark*	*Marginal mark*
Leave unchanged	··· under matter to remain	
Insert in text the matter indicated in the margin		New matter followed by or
Delete	through single character, rule or underline or through all characters to be deleted	or
Substitute character or substitute part of one or more word(s)	through letter or through characters	new character or new characters
Change to italics	under matter to be changed	
Change to capitals	under matter to be changed	
Change to small capitals	under matter to be changed	
Change to bold type	under matter to be changed	
Change to bold italic	under matter to be changed	
Change to lower case	Encircle matter to be changed	
Change italic to upright type	(As above)	
Change bold to non-bold type	(As above)	
Insert 'superior' character	through character or where required	or under character e.g. or
Insert 'inferior' character	(As above)	over character e.g.
Insert full stop	(As above)	
Insert comma	(As above)	
Insert single quotation marks	(As above)	or and/or or
Insert double quotation marks	(As above)	or and/or or
Insert hyphen	(As above)	
Start new paragraph		
No new paragraph		
Transpose		
Close up	linking characters	
Insert or substitute space between characters or words	through character or where required	
Reduce space between characters or words	between characters or words affected	