Content analysis-based documentation and exploration of research articles

Shwe Sin Phyo
Assistant Lecturer, Mandalay, Myanmar

Received 2 July 2020
Revised 18 December 2020
Accepted 1 January 2021

Abstract

Purpose – With the wealth of information available on the World Wide Web, it is difficult for anyone from a general user to the researcher to easily fulfill their information need. The main challenge is to categorize the documents systematically and also take into account more valuable data such as semantic information. The purpose of this paper is to develop a concept-based search system that leverages the external knowledge resources as the background knowledge for getting the accurate and efficient meaningful search results.

Design/methodology/approach – The paper introduces the approach which is based on formal concept analysis (FCA) with the semantic information to support the document management in information retrieval (IR). To describe the semantic information of the documents, the system uses the popular knowledge resources WordNet and Wikipedia. By using FCA, the system creates the concept lattice as the concept hierarchy of the document and proposes the navigation algorithm for retrieving the hierarchy based on the user query.

Findings – The semantic information of the document is based on the two external popular knowledge resources; the authors find that it will be more efficient to deal with the semantic mismatch problems of user need.

Originality/value – The navigation algorithm proposed in this research is applied to the scientific articles of the National Science Foundation (NSF). The proposed system can enhance the integration and exploration of the scientific articles for the advancement of the Scientific and Engineering Research Community.

Keywords Natural language processing, Semantic similarity, WordNet, Wikipedia, Formal concept analysis, Information retrieval

Paper type Research paper

1. Introduction

There is a growing number of available data in the World Wide web knowledge repository, knowledge is one of main assets for storing and retrieving information. The analysts discover the appropriate knowledge based on the knowledge gain of external sources (Hislop *et al.*, 2018; Ristoski and Paulheim, 2016). The issue of knowledge management is one of the challenges in several research development organizations. The analysts attempted too much attention to knowledge management and exploration (Van Erp *et al.*, 2017; Zhang *et al.*, 2017a; Lee *et al.*, 2008). Since human cognitive ability is limited, the big challenge is to detect and get the appropriate knowledge from and store it with human-friendly and effective expression.

In the research community, the number of scientific articles is being published day by day, and researchers face serious difficulties for exploring the articles to the areas of interest and

The author immeasurably appreciates and expresses deepest gratitude to her assessors: Dr. Myat Myat Min, Dr. Soe Yu Maw and Dr. Thin Mya Mya Swe for their encouragement, insightful comments and hard questions.

Foremost, the author would like to express her sincere gratitude to her supervisor Dr. Nyein Nyein Myo, Professor, Faculty of Information Science, University of Computer Studies, Mandalay. Her guidance helped me in all the time of research and writing of this research article. My completion of this article could not have been accomplished without the support of her.

The author would also like to thank her parents who supported her spiritually throughout my life. Any attempt at any level cannot be satisfactorily completed without the support of parents.

Finally, the author thanks all who in one way or another contributed in the completion of this article.

Data Technologies and Applications
Vol. 56 No. 1, 2022
pp. 24-43

2514-9288
DOI 10.1108/DTA-07-2020-0146

making sure that their research problems are certainly new. The previous searching of the articles mainly based on the similarity of the text characteristics. This process retrieves hundreds and thousands of articles for the query word, while most of them are unintended articles since they contain the text of the given query word. The researchers tried to develop literature management by using the citation data as the structured feature of the articles. These data contain the name, title and source of conference or journal publication, which are currently used to link each other. Without knowing these data, it would be quite time-consuming to get the desired articles (Du and Evans, 2011; Lu *et al.*, 2012; Raamkumar *et al.*, 2017; Pan *et al.*, 2018). It is challenging for the documentation of the articles with their semantic information. The main problem is how to automatically extract semantic information of the article and which mechanism is used in the development of scientific concept hierarchy.

The researchers utilized hand-built or external knowledge resources such as WordNet, Wikipedia, DBpedia or many other powerful data sources to capture the semantic understanding and solve the complexity of the semantic mismatch problems (Zhang *et al.*, 2017a; Ruiz-Martínez *et al.*, 2011; Al-Shboul and Myaeng, 2014; Ray *et al.*, 2010; Taieb *et al.*, 2014; Ezzikouri *et al.*, 2019; Liu *et al.*, 2012; Lopez-Arevalo *et al.*, 2017; Jiang *et al.*, 2017; Jiang *et al.*, 2015). Among them, Wikipedia and WordNet are commonly used to solve the problem of semantic mismatching. Wikipedia is the large- scale repository, and its link representation can be assumed as the text articles that contain encyclopedic information. WordNet is the lexical network and categorize the words according to their meanings. It is an ideal source of knowledge for the semantic relationship between words or concepts. If the semantic knowledge of article is explored with powerful external knowledge resources and use this knowledge in the development of scientific concept hierarchy, the framework can help the user to explore their interest information.

Another core activity is to create the concept hierarchy with the suitable knowledge representation mechanism. Among various knowledge representation mechanisms, formal concept analysis (FCA) offers a mathematical formalism of concepts that represents extension through a collection of formal objects (extent) and intension by a set of formal attributes (intent). It is also a way of evaluating and exploring data and can be utilized in a range of applications, such as medical science, linguistics, information science and so on. It is an unsupervised classification that provides an international description for clusters, which is used in lattice-based retrieval methods. In the information retrieval (IR) domain, the lattice generated by FCA has demonstrated in document indexing and navigation strategy (Davey and Priestley, 2002; Poelmans *et al.*, 2013; Codocedo and Napoli, 2015). The researchers also considered the semantic information to the formal concepts creation, while most of them are domain specific.

The contribution of the proposed work is finding the semantic representation of articles and visualizes them as the concept hierarchy for the different domains of scientific articles collection. Finally, the searching algorithm for the constructed concept hierarchy for handling the semantic mismatch problem between the user query and the documentation of scientific articles is proposed.

2. Research background

Knowledge is a vital resource to understand a given situation or event. The knowledge management procedure collects, analyses, manages and shares the relevant data across an organization. The continuous amount of data is rising tremendously that the procedure of knowledge management is becoming an important.

2.1 Knowledge management

The search engine is the most popular knowledge management tool such as web search engine, retrieval engine for scholarly literature and the organization's data and document

query framework. One primary method of knowledge manipulation is knowledge retrieval. With the growth of information technology, the search engine can now efficiently retrieve information. Instead of the logical relations of the information components, the search engine focused on the similarities and the strength of the connections (Antoun *et al.*, 2016). The PageRank algorithm is used extensively in search engines (Gleich and Saunders, 2009), and the dominant method is a fuzzy search (Casillas *et al.*, 2013). Since its words similarity, a significant volume of unrelated information is retrieved.

For this limitation, Google first suggested improving its semantic search engine. This search engine encodes the semantic details of the knowledge entities whose dependence and causality are considered to be different from the intensity of the physical link. This information is used in various knowledge management and knowledge bases applications, such as Wikidata, DBpedia, Freebase, YAGO, Knowledge Vault of Google and the entity graph of Facebook and so on (Ristoski and Paulheim, 2016; Van Erp *et al.*, 2017; Färber *et al.*, 2018). The development of document management in scientific knowledge mining and management is a significant process as the well-form citation and data. The connections among the scientific articles are used by their cited and cocited information. The scientific knowledge graph is created, and this graph is used for the connection of a specific topic (Zhang *et al.*, 2017b). The evaluation of the research trend is based on temporal citation frequency, the betweenness centrality, and the alternation of citations based on the time period is utilized to discover significant articles and evaluate the research trend (Biswas *et al.*, 2016).

The automatic tools are developed for the facilitation of bibliometric science mappings, such as VOSviewer, CiteSpace and HistCite (Pan *et al.*, 2018). The researchers considered the connection of the articles with the information of citation and reference sections (Zhang *et al.*, 2017b). The use of citation data for quantifying the connections of the scientific articles is objective and clear. It is also efficient to conduct the analysis of the research trends. The citation analysis can be used to the relations between the primary topics contained in the scientific articles together with keyword extraction. Nevertheless, but the citation data restrict the knowledge of the article's content, and the graph can be assumed as trivial knowledge graph.

To sum up, exploration and representation are essential tasks for knowledge management. The visualization of knowledge representation with a graph-based approach is widely used and easy to analyze. Since knowledge exploration (KE) remains an important research area, the exploration should be designed and trained specifically for various types and formats of the data source.

2.2 Knowledge acquisition with large-scales knowledge repository

Knowledge is a meaningful resource for the understanding of the truth or reality of nature. Knowledge resources are the repositories that provide an explicit representation of knowledge about one or more domains. In particular, the knowledge resources can be divided into three types: (1) Domain- dependent ontologies, (2) Wikipedia and (3) WordNet.

Previous works used the features of different knowledge resources for the various applications. For example, Ruiz-Martínez, Valencia-García and their colleagues applied the domain-dependent ontology UMLS (Unified Medical Language System concepts) to learn the semantic relation and association of the biomedical concepts for biomedical natural language documents (Ruiz-Martínez *et al.*, 2011).

In 2014, the analysts used the WordNet taxonomy structure to find the methods of semantic similarity between words (Taieb *et al.*, 2014). The WordNet database consists of four classes, nouns, verbs, adjectives and adverbs. These classes are grouped by their semantic meaning and their lexical relations (Miller *et al.*, 1988, 1990). The researchers used the WordNet for quantifying the semantic similarity of the words (Ezzikouri *et al.*, 2019; Liu *et al.*, 2012; Lopez-Arevalo *et al.*, 2017). In 2019, the authors combined synsets and glosses of

WordNet lexical dictionary with the concepts of set theory for quantifying the similarity of two concepts (Ezzikouri *et al.*, 2019). In 2012, the scientists defined the concept hierarchy tree and graph to construct the concept vector model using the concept of WordNet hierarchy information. In this model, the tree and graph of the hierarchy definition are used to measure the closeness of semantic similarity (Liu *et al.*, 2012). For solving the problem of word-sense disambiguation (WSD), Lopez and his colleagues selected the predominant synset of WordNet for the sports and finance domain. They also used a structured corpus such as the British National Corpus (BNC) to accomplish the task (Lopez-Arevalo *et al.*, 2017).

The authors employed Wikipedia and DBpedia for the implementation of metallic materials knowledge graph. First, they extracted metallic material entities from DBpedia. Because of the absence of materials properties in DBpedia, the system used the Wikipedia category for enriching their approach (Zhang *et al.*, 2017a). The other authors used the feature of WordNet and Wikipedia for supporting the semantic approach of question classification. First, WordNet is used for entity type of the given word or phrase and used the Wikipedia for the presence of entities related to the word or phrase in the question (Ray *et al.*, 2010). For the application of query expansion, the researchers also use the conceptual links structure of the Wikipedia knowledge graph to enrich the semantic information of the original queries (Al-Shboul and Myaeng, 2014). The above research works are example descriptions under the usage of the different knowledge resources.

3. Formal concept analysis (FCA)

FCA is a novel strategy for analyzing and visualizing data, in order to make them more understandable (Poelmans *et al.*, 2013). It is the formal representation of concept theory, and its order set representation is applied in the field of mathematics. Previous studies used FCA for semantic concept analysis (Ferré and Cellier, 2020), domain knowledge representation (Grissa *et al.*, 2020), portal retrieval engine construction (Negm *et al.*, 2017), clustering (Balasubramaniam, 2015), FCA notion with e-learning (Beydoun, 2009) and so on.

In 2009, the researcher used FCA to support learning process of a student. They constructed a semantic web with the collective meta-knowledge structure for guiding the students. As a knowledge acquisition tool, they used FCA to process the student's virtual browsing trails to express and manipulate the dependencies between web pages to generate subsequent and more reliable search results (Negm *et al.*, 2017). FCA theory is also used for the representation of domain knowledge and concept exploration (Balasubramaniam, 2015; Beydoun, 2009). The former is the discovery of knowledge of complex metabolomics data, combining supervised classifiers, pattern exploring and FCA. The latter is the framework of the retrieval engine that is based on the portal data. The scientists have applied the FCA for document clustering with the combination of the fuzzy set theory (Grissa *et al.*, 2020). When compared to the existing query models, their system is sufficient and easy to retrieve the documents. In 2019, the analysts applied the classical FCA theory in the knowledge graph construction of projected graph patterns (Ferré and Cellier, 2020). The above research works are the instance description under the usage of FCA theory with the combination of fuzzy set theory, pattern structure, logical and relational concept analysis, triadic and so on.

The analysts applied the FCA to analyze, extract, learn and explore data. The input data (formal context) are the tabular form of object–attribute relation. The rows stand for objects set, and the columns stand for attributes set. It is a Boolean value's representation, that is, an object has or does not have a particular attribute. The notation of formal concepts is based entirely upon formal context representation. The triple K (formal context) consists of a set of objects (extent), attributes (intent) and connected by their binary relationship, $K=(G, M, I)$, where G, M are the collections of objects and attributes, and I is the relation of the objects-attributes, also defined as $I \subseteq G \times M$. The notion $(g, m) \in I$ can be defined as "object g has

attribute m". Formal concept (A, B) can be defined as $A' = B$ and $B' = A$, where, $A \subseteq G$ and $B \subseteq M$.

$$A' = \{m \in M \mid \forall\, g \in A \ : g\ I\ m\} \tag{1}$$

$$B' = \{g \in G \mid \forall\ m \in B \ : g\ I\ m\} \tag{2}$$

It is a pair with a set of all attributes shared by all objects from A and dually a set of all objects sharing all attributes from B. The set of formal concepts is based on the above notion. The generated formal concepts are used for the visualization of formal concept lattice (concept hierarchy). The formal concepts are arranged by the partial order, $\preccurlyeq$ (sub-concept super-concept relation) such that for any two formal concepts (A_1, B_1) and (A_2, B_2), (A_1, B_1) $\preccurlyeq$ (A_2, B_2) iff ($A_1 \subseteq A_2$) (equivalently, $B_2 \subseteq B_1$). The concept lattice visualization is ordered by the rule of generalization/specialization (partial order) relation. The concept lattice is defined in its decrease form in which objects and attributes are placed in next to their object/attribute concept, i.e. the most general concept for attributes derived from higher to lower orders and the most specific concept for objects shared from lower to higher orders. FCA provides a graphic view of object–attribute relation that is easy to navigate and use. The formal concepts are the main core of graph visualization that must be defined before the graph is built. These formal concepts are used in lattice construction which represents the knowledge structure that is more informative than general tree-like conceptual structures (Davey and Priestley, 2002; Poelmans *et al.*, 2013; Codocedo and Napoli, 2015).

4. Concept hierarchy foundation

The semantic representation of the document is the foundation of the concept hierarchy development. NLP-based feature extraction and knowledge exploration are the two main phases for the semantic representation of the articles. The former one extracts the specific features of the article. Then, the semantic knowledge of each feature is defined by the valuable knowledge of WordNet and Wikipedia. The articles generally include six integral sections (citation, title, abstract, main body, conclusion and reference). Among them, the system uses the content of essential parts (title and abstract).

4.1 NLP-based feature extraction

In general, the terms in academic article can be defined as the features of each article. All terms of the article are not identified as key features of each article. Therefore, the system analyzes the important terms by utilizing NLP-based feature extraction.

Based on the study of computational linguistics, every complete sentence is composed of two parts: subject and predicate. The subject (noun, pronoun or noun phrase) can be defined as the main body of the sentence, and predicate indicates what the subject is or does. The verb can also be defined as an important part of the sentence because it is the "action" word that tells the listener or reader what is happening in the sentence. Thus, the subject and verb can be assumed as the key role of the sentence. Therefore, the system uses the common natural language processing task for extracting the key role of the sentence. Based on the dependency parse tree, the sentence tree is created for each sentence. Then the system detects the Part-of-Speech (POS) tagging of each word. By using the entity recognition with regular expression pattern, the sub-syntax trees are retained. The pure noun phrases, nouns and verbs are extracted in each sub-tree on the basis of the positions of verbs and prepositions. Then repetitive noun phrases, nouns and verbs are eliminated (Hirschberg and Manning, 2015; Chen *et al.*, 2016; Shen *et al.*, 2008). The collection of noun phrases and individual words (noun and verb) is applied in the consecutive steps of the KE process.

4.2 Knowledge exploration

In this section, the system defines the semantic information of each document. The semantic resources used for the effectiveness of IR have been reported in the literature (Ray *et al.*, 2010). The system used WordNet for the semantic information of single terms, and Wikipedia is used for compound noun phrase. In contrast to WordNet, Wikipedia is the description of numerous or things that consists of compound noun phrases. For the semantic information of noun and verb word, this system uses WordNet with corpus-based semantic similarity measure (Poorna and Ramkumar, 2018; Meymandpour and Davis, 2016). Wikipedia can be defined as the directed knowledge graph since it consists of the Wikipedia articles as the nodes and the links which are connected by the directed link structure. In graph theory, the importance of nodes can be defined by their in-degree and out-degree value (Yang *et al.*, 2017; Zager and Verghese, 2008). This assumption is applied in the system for getting the semantic information of the extracted noun phrases.

4.2.1 Knowledge exploration for single terms. In the WordNet database, most of the words consist of more than one synset (semantic concept). The system uses the WordNet semantic similarity measure for selecting the appropriate synset. The basic similarity measures of WordNet are information content (IC)-based measures and distance-based measures. IC is the information value carried by a term or a word within a context (piece of text, document or corpus). IC-based measure computes the semantic relation between any two words by examining their IC contained in the word pairs. In distance-based measure, the similarity of two words is based on their semantic distance.

IC-based measure considers the integration of WordNet conceptual network and corpus to improve the performance of similarity measure. Much more information between two synsets has been shared, much more semantic similarity has been contained and vice versa. Dekang Lin IC-based measure (Lin) is one of the IC-based measure that returns a similarity score of two synsets, based on IC of two input synsets and their LCS (Lowest Common Subsumer) (Lin, 1998). Equation (3) is the Lin similarity between two synsets.

$$\mathrm{sim}_{\mathrm{lin}}(s_1, s_2) = \frac{2\mathrm{IC}_{\mathrm{corpus}}(s_{\mathrm{lcs}})}{\mathrm{IC}_{\mathrm{corpus}}(s_1) + \mathrm{IC}_{\mathrm{corpus}}(s_2)} \tag{3}$$

where s_1 and s_2 are the WordNet synsets for the noun or verb word. $\mathrm{IC}_{\mathrm{corpus}}(s_1)$, $\mathrm{IC}_{\mathrm{corpus}}(s_1)$ and $\mathrm{IC}_{\mathrm{corpus}}(s_{\mathrm{lcs}})$ are the IC of WordNet synsets and their LCS. s_{lcs} is the most specific common parent of synset "s_1" and "s_2" from the root node. Note that for any IC measure, the result is dependent on the corpus used to generate the IC and the specifics of how IC was created. The Brown corpus is a general text collection that is commonly used in text linguistics, containing samples of 500 English-language text documents and approximately one million words. In this work, the IC of the synset is based on Brown corpus.

Table 1 shows three of WordNet synsets for the noun word "polymorphism". This representation is used as the semantic information of each noun or verb word. To get the most suitable synset, the system measures the semantic relation between each possible synsets

Word	Synset	Definition
Polymorphism	polymorphism.n.01	(genetics) the genetic variation within a population that natural selection can operate on
	polymorphism.n.02	(chemistry) the existence of different kinds of crystal of the same chemical compound
	polymorphism.n.03	(biology) the existence of two or more forms of individuals within the same animal species (independent of sex differences)

Table 1.
WordNet synsets of "polymorphism"

with the total synsets of the article.

$$\mathrm{sim}(w(s_k), w(s_{\mathrm{total}})) = \max\left(\mathrm{avg}\left(\sum \mathrm{lin}(w(s_k),\ w(s_{\mathrm{total}}))\right)\right) \tag{4}$$

In Equation (4), $w\ (s_{\mathrm{total}})$ is the total synsets of the article. $w\ (s_k)$ is the possible synsets of each noun or verb word. Since the similarity range is between "0" and "1", the system selects the synset with the maximum similarity value (Poorna and Ramkumar, 2018). The system finds their total similarity values and average. Finally, extracts the synset with the greatest similarity value. Since WordNet categories of noun and verb have different root nodes, their synsets selection process perform separately.

For instance, the semantic representation of "polymorphism" has three noun synsets. Thus, the system determines which synset is suitable for the word "polymorphism".

$$\mathrm{sim}(w_{\mathrm{polymorphism}}(s_k), w(s_{\mathrm{total}}) \leftarrow \max\left(\mathrm{avg}\left(\sum \mathrm{lin}(w_{\mathrm{polymorphism}}(s_k), w_N(s_{\mathrm{total}})\ )\right)\right) \tag{5}$$

$$\mathrm{lin}(w_{\mathrm{polymorphism}}(s_k), w_N(s_{\mathrm{total}})\ ) = \frac{(2 \times \mathrm{IC}_{\mathrm{corpus}} \mathrm{lcs}(w_{\mathrm{polymorphism}}(s_k), w_N(s_{\mathrm{total}})))}{(\mathrm{depth}(w_{\mathrm{polymorphism}}(s_k)) + \mathrm{depth}(w_N(s_{\mathrm{total}})))} \tag{6}$$

$w_N(s_{\mathrm{total}})$ is the total number of synsets for all noun words of the article. $w_{\mathrm{polymorphism}}(s_k)$ is each of possible synset for the noun word "polymorphism" (Table 1). By using the above Equations (5) and (6), the system finds the similarity value of three synsets with other synsets of remaining noun words. The appropriate synset is the synset in the pair that has the greatest similarity value. The appropriate synsets of the remaining noun or verb words are defined by using the above equations. The synset selection process of each article is performed separately, since they have different content (aim, method and so on).

4.2.2 Wikipedia-based knowledge exploration. The semantic representation of noun phrases is based on Wikipedia knowledge resource. In graph theory, a graph can be denoted as $G = (V, A)$, where "V" is the set of vertices and "A" is a collection of ordered pairs of vertices (the links between the vertices) or edges. This graph is also known as the directed graph. In general, the degree of a node is the total number of neighbors. A node is important if it has many neighbors that link to it or from it. However, the degree can be divided into in-degree (incoming links) and out-degree (outgoing links) in the direct graph. It can be defined by the following equations.

$$\deg^-(v) = |A|^- = \{a_i | (a_i, a)\ ,\ a \subseteq L\} \tag{7}$$

$$\deg^+(v) = |A|^+ = \{a_i | (a, a_i)\ ,\ a \subseteq L\} \tag{8}$$

The in-degree ($\deg^-(v)$) and out-degree ($\deg^+(v)$) values are the number of edges that come to the vertex and the number of edges going out of the vertex. When studying weighted networks and labeled node strength, the degree has been extended to the total of weights, so the weighted degree and the weighted in- and out- degree were determined. That help to define the critical nodes in a graph.

Wikipedia can be assumed as a directed graph $G\ (W_A, W_L)$, where W_A and W_L are Wikipedia articles and links. Each Wikipedia article $w_a \in W_A$ effectively summarizes its entity, i.e. the title of w_a, and provides links to the user to get the associated Wikipedia articles. The link structure of Wikipedia and the theory of a directed graph can be applied for defining the related links for a particular object (Zhang *et al.*, 2017a; Jiang *et al.*, 2017; Zager and Verghese, 2008). For example, the in-link and out-link of the noun phrases "knowledge base" are the total number of links in the Wikipedia and the set of links from the title and main body of the "knowledge base" page dump.

In-links (l_{in}) of article "w_a" is a set collection of all articles that indicate to the Wikipedia article "w_a". It can be written as

$$l_{in}\ (w_a) = \{w_{ai} | (w_{ai}, w_a) \in W_L\} \tag{9}$$

Out-links (l_{out}) of article "w_a" is a set collection of all links within the main body of the Wikipedia article "w_a".

$$l_{out}(\ w_a) = \{w_{ai}\ |(\ w_a, w_{ai}) \in W_L\} \tag{10}$$

4.2.3 Link selection. Link selection is based on associated in-links and out-links of each noun phrase. First, the system selects the set of links that include both in-link and out-link of the noun phrase. Then, find their weighted value and selects the important or relevant links for the noun phrase.

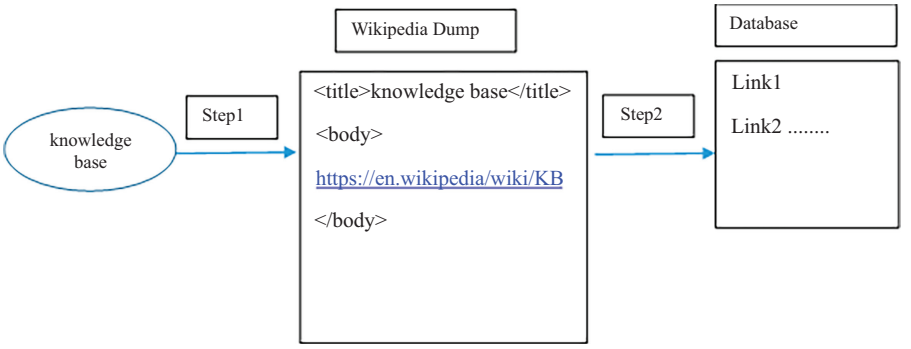

Figure 1.
In-link extraction for the noun phrase (knowledge base)

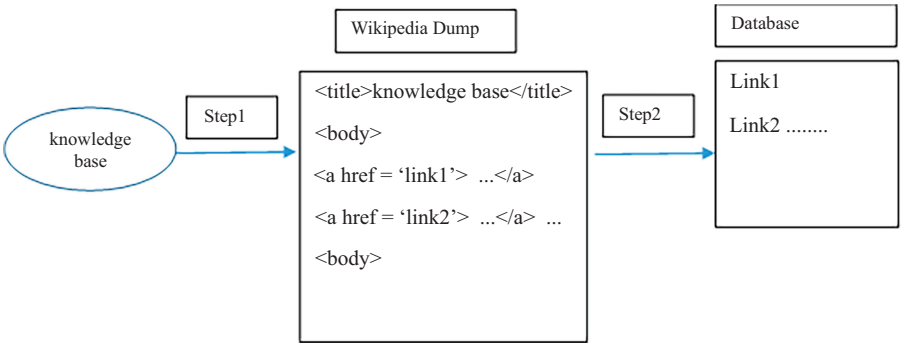

Figure 2.
Out-link extraction for the noun phrase (knowledge base)

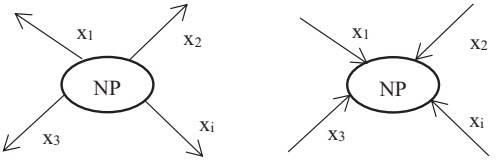

Figure 3.
In-link and out-link collection of noun phrase (knowledge base)

The in-links and out -links have been graphically presented in Figures 1–3. Equation (11) is the expansion links for each noun phrase based on the intersect operation of in-link and out-link collection. Then, the system calculates their in-degree and out-degree value. The top "k" degree value is selected and then extracted the links that include both in the top "k" of "*In*" and "*Out*" degree value set. The final Wikipedia links can be defined as the appropriate links for each noun phrase.

$$l_{np} = l_{in}\ (\mathrm{NP}) \cap\ l_{out}(\mathrm{NP}),\ \ l_{in}, l_{out} \in W_L \tag{11}$$

$$\deg^{-}(l_{np}) = |A|^{-} = \{a_i | (a_i, a)\ ,\ \ a \in W_L\} \tag{12}$$

$$\deg^{+}(l_{np}) = |A|^{+} = \{a_i | (a, a_i)\ ,\ \ a \in W_L\} \tag{13}$$

$$l_{np} = \mathrm{maxdeg}^{-}(l_{np}) \cap \mathrm{maxdeg}^{+}(l_{np}) \tag{14}$$

Table 2 represents the "*In*" and "*Out*" links of noun phrase and the extracted links collection. For the noun phrase ("Knowledge base"), the total number of in-links and out-links are 166 and 466. The intersection range of "*In*" and "*Out*" links is 60. For identifying the most related links of the noun phrase, their weighted score is used. Finally, the system selects the links based on top "k" weighted value, which serves as the appropriate information for the noun phrase.

5. Concept hierarchy development

The concept hierarchy of the system uses the data of NSF document library. It has a collection of awarded articles from 1960 to 2019. According to the NSF research area, there are seven main organizations for the awarded research articles. This section generates the concept hierarchy of NSF research articles by using their semantic information (Section 4).

5.1 Concept generation

Table 3 is the formal context (G, S, I) which is the binary relation (I) of article object collection (G) and the set of their semantic properties (S). The table data are randomly selected articles of Information and Intelligent Systems (IIS) and properties set "S". The object collection is the set of rows, and the properties are the set of columns. The link connection between them is illustrated in cross-sample "×", which means if the object "IIS1" has the semantic attribute "concept.n.01" then the relation between them is represented with cross-sample "×". If not, no value is added to the content of the binary table.

Within the context, there is a maximally general concept (supremum) comprising all of the articles as the extent and having an empty intent or semantic attributes. Furthermore, there is a minimally general concept (infimum) comprising no article at all and having all semantic attributes as the intent. The remaining concepts are created by using the following steps.

Table 2.
In-links and out-links of noun phrase (knowledge base)

Noun phrase	In-link length	In-links (l_{in})	Out-link length	Out-links (l_{out})	Common length	l(NP)
Knowledge base	166	Semantic reasoner, authority control, automated reasoning…	466	Knowledge representation and reasoning, semantic reasoner, logic programming…	60	Semantic reasoner, Yago (Database), library classification, authority control…

	agentive_role.n.01	concept.n.01	annotation.n.01	bourn.n.02	topology.n.04	correlate.v.02	fuse.v.03	stick_out.v.01	pixel.n.01	Knowledge discovery	Community structure	Data integration	Big data	Knowledge base
IIS1	×		×					×		×		×	×	×
IIS2		×	×					×	×			×		×
IIS3			×								×		×	
IIS4	×	×	×							×				×
IIS5			×											×
IIS6							×			×				
IIS7	×											×		
IIS8				×				×				×		
IIS9	×										×			
IIS10		×											×	×
IIS11						×	×							
IIS12					×	×						×		
IIS13							×	×						
IIS14		×			×	×					×			
IIS15	×				×									

Table 3.
Formal context

(1) Pick any article or articles set, A,

(2) Derive the semantic attributes A',

(3) Derive $(A')'$ and

(4) Finally, (A'', A') is a formal concept.

A dual approach can be taken starting with an attribute. For instance, take the attributes that belong to the article "IIS5" in Table 3. Then, collect the articles that have those attributes. The result is the set of article $D \subseteq G$, consisting of articles "IIS1", "IIS2", "IIS4" and "IIS5". The set "D" of article objects collection has the binary relation of semantic attributes set $A \subseteq S$, "knowledge base" and "annotation.n.01". Table 4 presents the formal concepts of Table 3.

Each article-semantic attribute pair is defined as the formal concept description of the given formal context. The set $A = D'$ is called the extent (articles), and the set $D = A'$ is called the intent (semantic attributes) of the formal pair (D, A).

Concepts	Formal concepts (extents and intents) set for corresponding notation
C0	('agentive_role.n.01', 'concept.n.01', 'annotation.n.01', 'stick_out.v.01', 'knowledge discovery', 'data integration', 'big data', 'knowledge base', 'pixel.n.01', 'community structure', 'fuse.v.03', 'bourn.n.02', 'correlate.v.02', 'topology.n.04') ()
C1	('agentive_role.n.01', 'concept.n.01', 'annotation.n.01', 'stick_out.v.01', 'knowledge discovery', 'data integration', 'big data', 'knowledge base') ('IIS1',)
C2	('annotation.n.01', 'stick_out.v.01', 'data integration', 'knowledge base', 'pixel.n.01') ('IIS2',)
C3	('concept.n.01', 'annotation.n.01', 'big data', 'community structure') ('IIS3',)
C4	('knowledge discovery', 'fuse.v.03') ('IIS6',)
C5	('stick_out.v.01', 'data integration', 'bourn.n.02') ('IIS8',)
C6	('agentive_role.n.01', 'community structure') ('IIS9',)
C7	('fuse.v.03', 'correlate.v.02') ('IIS11',)
C8	('stick_out.v.01', 'community structure', 'fuse.v.03') ('IIS13',)
C9	('concept.n.01', 'community structure', 'correlate.v.02', 'topology.n.04') ('IIS14',)
C10	('agentive_role.n.01', 'correlate.v.02') ('IIS15',)
C11	('annotation.n.01', 'stick_out.v.01', 'data integration', 'knowledge base') ('IIS1', 'IIS2')
C12	('concept.n.01', 'annotation.n.01', 'big data') ('IIS1', 'IIS3')
C13	('agentive_role.n.01', 'annotation.n.01', 'knowledge discovery', 'knowledge base') ('IIS1', 'IIS4')
C14	('agentive_role.n.01', 'data integration') ('IIS1', 'IIS7')
C15	('concept.n.01', 'big data', 'knowledge base') ('IIS1', 'IIS10')
C16	('concept.n.01', 'community structure') ('IIS3', 'IIS14')
C17	('correlate.v.02', 'topology.n.04') ('IIS12', 'IIS14')
C18	('stick_out.v.01', 'data integration') ('IIS1', 'IIS2', 'IIS8')
C19	('concept.n.01', 'big data') ('IIS1', 'IIS3', 'IIS10')
C20	('knowledge discovery',) ('IIS1', 'IIS4', 'IIS6')
C21	('fuse.v.03',) ('IIS6', 'IIS11', 'IIS13')
C22	('annotation.n.01', 'knowledge base') ('IIS1', 'IIS2', 'IIS4', 'IIS5')
C23	('data integration',) ('IIS1', 'IIS2', 'IIS7', 'IIS8')
C24	('stick_out.v.01',) ('IIS1', 'IIS2', 'IIS8', 'IIS13')
C25	('concept.n.01',) ('IIS1', 'IIS3', 'IIS10', 'IIS14')
C26	('community structure',) ('IIS3', 'IIS9', 'IIS13', 'IIS14')
C27	('correlate.v.02',) ('IIS11', 'IIS12', 'IIS14', 'IIS15')
C28	('annotation.n.01',) ('IIS1', 'IIS2', 'IIS3', 'IIS4', 'IIS5')
C29	('knowledge base',) ('IIS1', 'IIS2', 'IIS4', 'IIS5', 'IIS10')
C30	('agentive_role.n.01',) ('IIS1', 'IIS4', 'IIS7', 'IIS9', 'IIS15')
C31	() ('IIS1', 'IIS2', 'IIS3', 'IIS4', 'IIS5', 'IIS6', 'IIS7', 'IIS8', 'IIS9', 'IIS10', 'IIS11', 'IIS12', 'IIS13', 'IIS14', 'IIS15')

Table 4.
Concepts notation

Table 5 is a set of expansion links for the noun phrases, where each link is strongly connected to an individual semantic attribute or phrase. These links are assigned using the in-link and out-link score (section 4). For example, the noun phrase "data integration" is frequently associated with the expansion links, "data security, data compression, data mining and big data." Here, searching the query phrase "data integration" might work well for the queries such as "data security," "data compression," "data mining" and "big data." Instead of correlated to several individual query terms, the system will be correlated to the phrase query as a whole for solving the language ambiguity problem (see Table 5).

5.2 Concept hierarchy

The concept hierarchy is based on the formal model of conceptual data analysis and knowledge processing. By using this model, the system provides the graphical representation of articles and their (semantic) attributes. Every concept has a collection of all articles that belong to common semantic attributes. The relation between them is the partial order. Concept $C_2(D_2, A_2)$ is a super-concept of concept $C_1(D_1, A_1)$, since the extent of C_1 is a subset of the extent of C_2 and the intent of C_1 is a superset of the intent of C_2. The partial relation of formal concepts makes the concept hierarchy, and it represents the directed acyclic graph of concepts for the relation between a pair of objects and properties such that the objects share precisely their properties and these properties apply to precisely the objects. Consider the formal concept C8 ('stick_out.v.01', 'community structure', 'fuse.v.03') ('IIS13',) and C26 ('community structure',) ('IIS3', 'IIS9', 'IIS13', 'IIS14'). Since the extent of C8 is a subset of the extent of C26 and the intent of C8 is a superset of the intent of C26. Therefore, C8 is drawn below C26 and connected with a line. The ordered set of all formal concepts in (G, S, I) is denoted by $\mathcal{B}(G, S, I)$ and is called concept lattice (concept hierarchy).

Finally, the concepts are ordered by extent set-inclusion (or, dually, by intent set-inclusion), which is the formation of concept hierarchy for the research articles. This (partial) order makes the formal concept hierarchy have the supremum concept at the top, the infimum concept at the bottom and the other concepts between them. The concepts are linked upward (more general concepts and super-concepts) and downward (less general concepts and sub-concepts).

6. Semantic search: querying

The semantic search of the system relies on two steps, "precise" and the "partial". The former one is finding the most general concept where a user query is precisely related to the "query term" (indicated μ (ss) in (Davey and Priestley, 2002)). The result is the list of articles which is directly associated with the query term.

The second step is finding the relevant concepts for the "query term". Since the concepts are generated by using the semantic attributes and their related objects, the two concepts are relevant if they have same lower bound (successor) or sub-lower bound (Messai *et al.*, 2005). It can be assumed as the two concepts are relevant if they have one or more same semantic attributes or objects. Therefore, instead of extracting the all concepts subsumed by both

Noun phrase	Expansion links
Knowledge discovery	Internet, database and World Wide Web
Community structure	Computer network, social network, artificial neural network and network science
Knowledge base	Database, semantic web and prolog
Data integration	Data security, data compression, data mining and big data
Big data	Statistics and machine learning

Table 5.
Expansion links

concepts and them finding their super-concepts, the system uses the incremental structure of constructed concept hierarchy as the beneficial for the consideration of the relevancy relation. The formal concept creation of the system is based on the semantic set inclusion of objects and attributes. This means that two concepts (A_1, B_1) and (A_2, B_2) are relevant if there exists $(A_1 \cap A_2)$ or $(B_1 \cap B_2)$. The system determined the relevancy rank of two concepts by their number of equivalent extent or object because the user query is the intent or attribute. The precise concept (A_q, B_q) for query "q", the relevant concepts can be extracted by the following search algorithm.

(1) Find the precise concept (A_q, B_q) for the query $\{q\}$ in the lattice, $L(D, S, I)$ // D = documents, S = semantic attributes, I = incident relation

(2) Find the relevant concepts of (A_q, B_q) in L

- Create the semantic attribute set $(Q_s \subseteq S)$, which describes the extent of precise concept (A_q, B_q).
- Extract all concepts (A_r, B_r) in the concept hierarchy which describe the attributes or attribute in the Q_s.

(3) Find the relevancy value of each relevant concept, $c_r \subseteq (A_r, B_r)$.

$$\text{rank}(c_r, (A_q, B_q)) = 100 * \frac{|c_r.\text{extent} \cap (A_q, B_q).\text{extent}|}{|c_r.\text{extent}|} \tag{15}$$

Extract the relevant concepts, $\text{rank}(c_r, (A_q, B_q)) \geq 50\%$.

The result is composed of the extent (articles) in the precise and relevant concepts.

6.1 Query manipulation

Query manipulation allows getting more accurate results of user need. First, the user query is accepted, and the system considers the embedded meaning of the query. A query may be phrases or individual words. If the query is a single term, the possible semantic representation of the given query term is returned back to the user. Based on user confirmation, the system finds the articles of precise and relevant results. Query manipulation of phrase is finding the phrase query in the concept lattice and also considers the related links connection. The system extracts all articles of given phrase query and also the articles of related links connection. If the given phrase is not included in the concept lattice, the system determines the embedded meaning of each given term by using the WordNet semantic similarity measure.

For the user query "java drink", the system finds the articles that include the phrase "java drink". If there is no article that includes the "java drink" phrase, then the system converts the phrase to unigram words assumption. Finally, the system uses the WordNet semantic similarity measure (section 4) for the appropriate WordNet concepts and finds the articles.

6.2 Querying: example

Querying example is based on the concept lattice of Figure 4. Manipulate the user query "concept" to the semantic attribute "concept.n.01". Find the precise concept (A_q, B_q) of the query {concept.n.01}. The result is C25 (A_q { 'IIS1', 'IIS3', 'IIS10', 'IIS14'}, B_q {concept.n.01}). Create the semantic attributes set Q_s, ('agentive_role.n.01', 'concept.n.01', 'annotation.n.01', 'stick_out.v.01', 'knowledge discovery', 'data integration', 'big data', 'knowledge base', 'community structure''correlate.v.02', 'topology.n.04') which possess the extent of precise concept. Then the system finds the relevant concepts which include the attributes or attribute

in the Q_s. There are 11 relevant concepts for the semantic attribute "concept.n.01". The system calculates the rank value of each relevant concept by using Equation (15).

The rank result of the relevant concepts is based on the ratio of retrieved articles and their corrected articles. If the total number of articles is two and the corrected article is one then the ranking percentage can be assumed as 50%. To cover the relevancy relation, the system extracts the concepts with the rank percentage 50 and above. The following concepts can be assumed as the relevant concepts for the query concept.

(1) C17 {('IIS12', 'IIS14') ('correlate.v.02', 'topology.n.04')}

(2) C18 {('IIS1', 'IIS3', 'IIS10') ('concept.n.01', 'big data')}

(3) C26 {('IIS3', 'IIS9', 'IIS13', 'IIS14') ('community structure',)}

The precise concept includes the extent ('IIS1', 'IIS3', 'IIS10', 'IIS14'). The extent of C17 includes two and the corrected is one; therefore, the rank percentage is 50%. For concept C18, the total articles are three and the corrected is three; therefore, the rank percentage is 100%. For concept C26, the total articles are four and the corrected is two; therefore, the rank percentage is 50%. For the query word "concept", the result articles are ['IIS1', 'IIS3', 'IIS10', 'IIS14', 'IIS9', 'IIS13', 'IIS12'] which are the articles of precise concept (C25) and relevant concepts (C17, C18, C26).

7. Validation

The previous section defined the querying example with a semantic attribute concept.n.01. The result is ('IIS1', 'IIS3', 'IIS10', 'IIS14', 'IIS9', 'IIS13', 'IIS12') articles which is the set of articles that is directly and closely related to the semantic attribute. The testing of the querying result is based on out-of-sample testing of cross-validation. It is the statistical

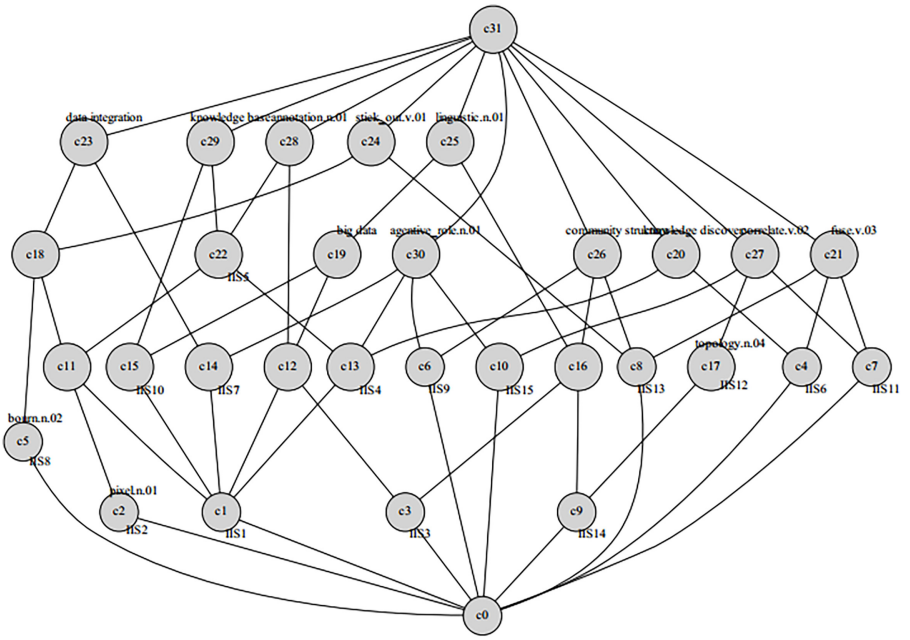

Figure 4.
Concept hierarchy

method for evaluating and comparing the learning algorithm by splitting a collection of data into two subgroups: one is used for analysis (training) and the other is used to validate (validation). In typical cross-validation, the training and validation must cross-over in successive rounds such that each data point has a chance of being validated against.

To access the performance of classification algorithms, both k-fold and leave-p-out cross-validation are very popular. Leave-P-Out cross-validation method leaves $'p'$ data points out of the training data, i.e. if the original sample has $'n'$ data points then $'n\text{-}p'$ samples are used to train the model and $'p'$ points are the validation. The original sample can be split in this way and then the error is calculated to provide total effectiveness for all the trials. A particular case of this approach is $p = 1$. It is regarded as the Leave–one-out cross-validation (LOOCV). This approach is usually chosen because it does not suffer from the intensive computation, as the number of potential combinations is equal to the number of data points in the original sample or $'n'$. For example, there are 100 samples, LOOCV will construct a model for each one of them (using the other 99 samples) and then calculate the mean (Refaeilzadeh *et al.*, 2009; Browne, 2000).

For the result of the semantic attribute "concept.n.01", the precise articles ('IIS1', 'IIS3', 'IIS10', 'IIS14') can be included in the relevant articles ('IIS12', 'IIS14', 'IIS1', 'IIS3', 'IIS10', 'IIS9', 'IIS13', 'IIS14'). As for the precise articles, the system checks whether the proposed method can find them with the modification of the formal context (remove the binary relation of semantic attribute "concept.n.01" and the directly related articles, 'IIS1', 'IIS3', 'IIS10', 'IIS14'). For instance, the binary relation of "IIS1" and "concept.n.01" is removed in the original formal context and train this context for getting the reformed concept lattice. In this situation, the system cannot retrieve "IIS1" as the directly related article, but it can be included in the set of relevant articles.

As for the relevant articles, the system checks how it evolves after applying the proposed searching approach to the reformed concept lattice, as the removal of a (article, semantic attribute) binary relation can affect the structure of the concept lattice and also the result of the proposed searching approach. The total effectiveness of the system can be illustrated by small differences in the content of this set with the adapted LOOCV approach.

8. Evaluation

The evaluation is based on whether the system can find the extent if the system eliminates the relation between the extent and intent individually. It is the intentionally eliminating a single (article, semantic attribute) relation from the original formal context and reformed the concept lattice (C_{re}). The pair semantic attribute (s) and the article (IIS) is used for identifying the reformed lattice or reformation (C_{re}) and the relation of which was removed for the creation of C_{re}.

From the C_{re}, if the eliminated article can be retrieved by querying for the attribute, then the C_{re} is efficient. (querying the "concept.n.01" from C_{re} and the article "IIS1" can retrieve from C_{re}). It should be noted that the total number of C_{re} for a query is defined by the total number of directly related articles because only a single (article, semantic attribute) relation is removed from the formal context at each time (e.g. for the "concept.n.01", the system constructs four C_{re} for articles 'IIS1', 'IIS3', 'IIS10', 'IIS14'). The following equation is the efficiency rate for a given query (s), where, total $C_{re}(s)$ refers to the total number of $C_{re}(s)$ developed using query "s" and the set of precise articles. efficient $C_{re}(s)$ is total number of efficiency rate that can be assumed as an achievement.

$$\text{efficiency rate} = \frac{\text{efficient } C_{re}(s)}{\text{total } C_{re}(s)} \tag{16}$$

The value of efficiency rate indicates that the article can be retrieved for a query "s" even if the article and this query is not directly related. Equations (17)–(19) are the calculation of

precision, recall and *f*-measure. The precision is based on the proportion of true positives over the retrieved lists of articles in the reformed concept hierarchy. The recall value is the proportion of true positives over the retrieved lists of articles in the original concept hierarchy (Browne, 2000; Manning *et al.*, 2008). $\mathrm{Ret}(\mathrm{modified}_{\mathrm{artlcies}}, s)$ is the total articles retrieved from C_{re} (precise ∪ relevant) querying for the semantic attribute"*s*," and the $\mathrm{Ret}(\mathrm{original}_{\mathrm{articles}}, s)$ is the original concept hierarchy. Then, *f*-measure is calculated for defining the weighted average of precision and recall.

$$\mathrm{precision}\ (\mathrm{modified}_{\mathrm{artlcies}}, s) = \frac{\mathrm{Ret}(\mathrm{modified}_{\mathrm{artlcies}}, s) \cap \mathrm{Ret}(\mathrm{original}_{\mathrm{articles}}, s)}{\mathrm{Ret}(\mathrm{modified}_{\mathrm{artlcies}}\ , s)} \tag{17}$$

$$\mathrm{recall}(\mathrm{modified}_{\mathrm{artlcies}}, s) = \frac{\mathrm{Ret}(\mathrm{modified}_{\mathrm{artlcies}}, s) \cap \mathrm{Ret}(\mathrm{original}_{\mathrm{articles}}, s)}{\mathrm{Ret}(\mathrm{original}_{\mathrm{articles}}, s)} \tag{18}$$

$$F - \mathrm{Measure} = 2^{*}\frac{(\mathrm{Recall}^{*}\mathrm{Precision})}{(\mathrm{Recall} + \mathrm{Precision})} \tag{19}$$

The precision, recall and *f*-measure are calculated for each semantic attribute with their C_{re}. The test set data are the awarded articles of NSF (2018) that under in the divisions of IIS, biological infrastructure (BI), environmental biology (EB),earth sciences (ES), which include 545 records, 138 records, 220 records and 1052 records. From the above tested datasets of 1955 records, the system selected 150 records for each division (about 600 records). The randomly selected queries words are about 250 for all divisions and reformed the formal context almost 1600 times.

Table 6 is the results of the measure for ten sample queries. For the query "anti-racketeering law", the efficiency rate of 1 means that all reformation (C_{re}) are efficient. Recall and precision of 0.7, 0.9 mean that for the elimination of 10% of the relations for a semantic attribute (1 over 10 articles), still 70% of information were retrieved and 90% of the information were correct.

It is worth noticing that a positive relationship exists between the number of articles in which the semantic attribute occurs and measure of the efficiency rate. The semantic attribute occurring in a few articles would be in fewer concepts in the concept hierarchy; therefore, the simulation affects them in the worst manner. For instance, the elimination of one relationship with an article leads to the elimination of 25% of its relationships (1/4) for a semantic attribute "fauna". For the "curium", the elimination of one relationship with an article leads to the elimination of only 50% of its relations (1/2).

Table 7 shows the positive predicted value and the number of correct positive predictions for the test set data. In a wider sense, precision and recall maintain their values above 80% over all the tested samples. It should be noted that semantic retrieval is the basic objective of proposed algorithm. The retrieved articles is dependent on the order in which the algorithm is presented, and the order can also depend on user preferences and allows to change a threshold (rank value) for the final articles set.

It should also be noted that a certain degree of bias is to be expected and is likely to be caused by the inclusion of the directly related articles in the precision/recall measures. In other words, since for each C_{re}, the system eliminates only one (article, semantic attribute) relation, the rest of directly related articles will be include in both sets retrieved when querying the C_{re} and the original concept hierarchy. Therefore, the precision/recall measures are intended to be used as a mean of examining how the set of relevant concepts is affected for each semantic attribute should not be viewed as a medium of contrast with other approaches to IR.

Finally, even though more experiments have to be carried out, the relevant concept consideration of the system is valuable and can be used for the exploration of concept lattice as a semantic concept hierarchy to retrieve objects that are not directly related to a query.

Table 6.
Evaluation measures over articles per query

Query	Number of articles Precise	Relevant	Precision	Recall	Efficiency rate
Fauna	4	19	1	0.75	0.72
Curium	2	11	1	0.5	0.18
Marsh	5	28	0.79	0.8	0.8
Agentive role	5	13	0.76	0.9	0.67
Anti-racketeering law	10	36	0.9	0.7	1
Metamorphosis	9	34	1	0.89	1
Monetary value	24	31	0.79	0.96	0.95
Computerized tomography	3	5	1	0.67	0
Biomass	5	6	1	0.83	0.66
Implosion therapy	3	4	0.89	0.67	0.67

Table 7.
The evaluation results

Division	Precision	Recall	*F*-measure
Information and Intelligent Systems (IIS)	0.85	0.89	0.86954
Biological infrastructure (BI)	0.9	0.91	0.9
Environmental biology (EB)	0.88	0.9	0.8898
Earth sciences (ES)	0.83	0.87	0.8495

9. Conclusion

This paper presents the conceptual hierarchy of research articles by using their semantic information of abstract and title. The impressive semantic characteristics of WordNet conceptual network and large-scale knowledge repository of Wikipedia are utilized for the semantic information about articles. According to the nature of different resources, recognition of semantic information is based on semantic similarity measure and link extraction method. This information is necessary to complete the full content of an article and is extremely valuable. The system uses the resulting semantic information into the classical FCA theory for generating semantic concepts of research articles. Finally, the semantic lattice is established and that lattice is used as the semantic concept hierarchy. This graph can support the semantic ambiguity and KE of research articles and can help to gain a deeper understanding of the article and thus to provide the user with richer retrieval results with regard to the content of research articles.

In future work, many sources could be used to obtain additional information about documents. For example, a comprehensive multi-domain ontology, DBpedia that provides Wikipedia content with semantics. Based on DBpedia information, the document is categorized by their particular category, and this information can be used to enrich the semantic representation of the document. Using this information, the FCA can be extended to relational concept analysis (RCA), which allows the relationships between objects to be taken into account in the FCA theory. Therefore, the lattice connects both the content and categories of the documents. From this lattice, the documents can be queried with the RCA system, and these documents are categories under the categories of comprehensive multi-domain ontology. The representation of concept hierarchy can also be visualized by the combination of classical FCA theory with other appropriate set theories. Another perspective is that the link extraction process can be extended with several methods of graph theory. WordNet synsets can also be explored with different similarity methods for different domain areas.

References

Al-Shboul, B. and Myaeng, S.H. (2014), "Wikipedia-based query phrase expansion in patent class search", *Information Retrieval*, Vol. 17 Nos 5-6, pp. 430-451.

Antoun, C., Zhang, C., Conrad, F.G. and Schober, M.F. (2016), "Comparisons of online recruitment strategies for convenience samples: Craigslist, Google AdWords, Facebook, and Amazon Mechanical Turk", *Field Methods*, Vol. 28 No. 3, pp. 231-246.

Balasubramaniam, K. (2015), "Hybrid fuzzy-ontology design using FCA based clustering for information retrieval in semantic web", *Procedia Computer Science*, Vol. 50, pp. 135-142.

Beydoun, G. (2009), "Formal concept analysis for an e-learning semantic web", *Expert Systems with Applications*, Vol. 36 No. 8, pp. 10952-10961.

Biswas, S., Sengupta, D., Bhattacharjee, R. and Handique, M. (2016), "Text manipulation using regular expression", in *2016 IEEE 6th International Conference on Advanced Computing (IACC)*, IEEE, pp. 62-67.

Browne, M.W. (2000), "Cross-validation methods", *Journal of Mathematical Psychology*, Vol. 44 No. 1, pp. 108-132.

(2013), in Casillas, J., Cordón, O., Triguero, F.H. and Magdalena, L. (Eds), *Interpretability Issues in Fuzzy Modeling*, Springer, Vol. 128.

Chen, W., Zhang, M., Zhang, Y. and Duan, X. (2016), "Exploiting meta features for dependency parsing and part-of-speech tagging", *Artificial Intelligence*, Vol. 230, pp. 173-191.

Codocedo, V. and Napoli, A. (2015), "Formal concept analysis and information retrieval–a survey", in *International Conference on Formal Concept Analysis*, Springer, Cham, pp. 61-77.

Davey, B.A. and Priestley, H.A. (2002), *Introduction to Lattices and Order*, 2nd ed., Cambridge University Press, New York.

Du, J.T. and Evans, N. (2011), "Academic users' information searching on research topics: characteristics of research tasks and search strategies", *The Journal of Academic Librarianship*, Vol. 37 No. 4, pp. 299-306.

Ezzikouri, H., Madani, Y., Erritali, M. and Oukessou, M. (2019), "A new approach for calculating semantic similarity between words using WordNet and set theory", *Procedia Computer Science*, Vol. 151, pp. 1261-1265.

Färber, M., Bartscherer, F., Menne, C. and Rettinger, A. (2018), "Linked data quality of dbpedia, freebase, opencyc, wikidata, and yago", *Semantic Web*, Vol. 9 No. 1, pp. 77-129.

Ferré, S. and Cellier, P. (2020), "Graph-FCA: an extension of formal concept analysis to knowledge graphs", *Discrete Applied Mathematics*, Vol. 273, pp. 81-102.

Gleich, D.F. and Saunders, M. (2009), *Models and Algorithms for Page Rank Sensitivity*, Stanford University, Stanford, CA.

Grissa, D., Comte, B., Petera, M., Pujos-Guillot, E. and Napoli, A. (2020), "A hybrid and exploratory approach to knowledge discovery in metabolomic data", *Discrete Applied Mathematics*, Vol. 273, pp. 103-116.

Hirschberg, J. and Manning, C.D. (2015), "Advances in natural language processing", *Science*, Vol. 349 No. 6245, pp. 261-266.

Hislop, D., Bosua, R. and Helms, R. (2018), *Knowledge Management in Organizations: A Critical Introduction*, Oxford University Press.

Jiang, Y., Zhang, X., Tang, Y. and Nie, R. (2015), "Feature-based approaches to semantic similarity assessment of concepts using Wikipedia", *Information Processing and Management*, Vol. 51 No. 3, pp. 215-234.

Jiang, Y., Bai, W., Zhang, X. and Hu, J. (2017), "Wikipedia-based information content and semantic similarity computation", *Information Processing and Management*, Vol. 53 No. 1, pp. 248-265.

Lee, S., Huh, S.Y. and McNiel, R.D. (2008), "Automatic generation of concept hierarchies using WordNet", *Expert Systems with Applications*, Vol. 35 No. 3, pp. 1132-1144.

Lin, D. (1998), "An information-theoretic definition of similarity", in Icml, Vol. 98 No. 1998, pp. 296-304.

Liu, H., Bao, H. and Xu, D. (2012), "Concept vector for semantic similarity and relatedness based on WordNet structure", *Journal of Systems and Software*, Vol. 85 No. 2, pp. 370-381.

Lopez-Arevalo, I., Sosa-Sosa, V.J., Rojas-Lopez, F. and Tello-Leal, E. (2017), "Improving selection of synsets from WordNet for domain-specific word sense disambiguation", *Computer Speech and Language*, Vol. 41, pp. 128-145.

Lu, M., Bangalore, S., Cormode, G., Hadjieleftheriou, M. and Srivastava, D. (2012), "A dataset search engine for the research document corpus", in *2012 IEEE 28th International Conference on Data Engineering*, IEEE, pp. 1237-1240.

Manning, C.D., Raghavan, P. and Schutze, H. (2008), *Introduction to Infor-Mation Retrieval?*, Cambridge University Press, pp. 405-416, Ch, 20.

Messai, N., Devignes, M.D., Napoli, A. and Smaïl-Tabbone, M. (2005), "Querying a bioinformatic data sources registry with concept lattices", in *International Conference on Conceptual Structures*, Springer, Berlin, Heidelberg, pp. 323-336.

Meymandpour, R. and Davis, J.G. (2016), "A semantic similarity measure for linked data: an information content-based approach", *Knowledge-Based Systems*, Vol. 109, pp. 276-293.

Miller, G., Fellbaum, C., Kegl, J. and Miller, K. (1988), "Wordnet: an electronic lexical reference system based on theories of lexical memory", *Revue Quebecoise de Linguistique*, Vol. 17 No. 2, pp. 181-212.

Miller, G.A., Beckwith, R., Fellbaum, C., Gross, D. and Miller, K.J. (1990), "Introduction to WordNet: an on-line lexical database", *International Journal of Lexicography*, Vol. 3 No. 4, pp. 235-244.

Negm, E., AbdelRahman, S. and Bahgat, R. (2017), "PREFCA: a portal retrieval engine based on formal concept analysis", *Information Processing and Management*, Vol. 53 No. 1, pp. 203-222.

Pan, X., Yan, E., Cui, M. and Hua, W. (2018), "Examining the usage, citation, and diffusion patterns of bibliometric mapping software: a comparative study of three tools", *Journal of Informetrics*, Vol. 12 No. 2, pp. 481-493.

Poelmans, J., Kuznetsov, S.O., Ignatov, D.I. and Dedene, G. (2013), "Formal concept analysis in knowledge processing: a survey on models and techniques", *Expert Systems with Applications*, Vol. 40 No. 16, pp. 6601-6623.

Poorna, B. and Ramkumar, A.S. (2018), "Semantic similarity measures: an overview and comparison", *International Journal of Advanced Research in Computer Science*, Vol. 9, Special Issue 1, p. 100.

Raamkumar, A.S., Foo, S. and Pang, N. (2017), "Using author-specified keywords in building an initial reading list of research papers in scientific paper retrieval and recommender systems", *Information Processing and Management*, Vol. 53 No. 3, pp. 577-594.

Ray, S.K., Singh, S. and Joshi, B.P. (2010), "A semantic approach for question classification using WordNet and Wikipedia", *Pattern Recognition Letters*, Vol. 31 No. 13, pp. 1935-1943.

Refaeilzadeh, P., Tang, L. and Liu, H. (2009), "Cross-validation", *Encyclopedia of Database Systems*, Vol. 5, pp. 532-538.

Ristoski, P. and Paulheim, H. (2016), "Semantic web in data mining and knowledge discovery: a comprehensive survey", *Journal of Web Semantics*, Vol. 36, pp. 1-22.

Ruiz-Martínez, J.M., Valencia-García, R., Fernández-Breis, J.T., García-Sánchez, F. and Martínez-Béjar, R. (2011), "Ontology learning from biomedical natural language documents using UMLS", *Expert Systems with Applications*, Vol. 38 No. 10, pp. 12365-12378.

Shen, L., Champollion, L. and Joshi, A.K. (2008), "LTAG-spinal and the treebank", *Language Resources and Evaluation*, Vol. 42 No. 1, pp. 1-19.

Taieb, M.A.H., Aouicha, M.B. and Hamadou, A.B. (2014), "A new semantic relatedness measurement using WordNet features", *Knowledge and Information Systems*, Vol. 41 No. 2, pp. 467-497.

Van Erp, M., Hellmann, S., McCrae, J.P., Chiarcos, C., Choi, K.S., Gracia, J., Hayashi, Y., Koide, S., Mendes, P., Paulheim, H. and Takeda, H. (Eds) (2017), *Knowledge Graphs and Language*

Technology: ISWC 2016 International Workshops: KEKI and NLP and DBpedia, Kobe, Japan, October 17-21, 2016, Vol. 10579, Springer, revised selected papers.

Yang, Y., Xie, G. and Xie, J. (2017), "Mining important nodes in directed weighted complex networks", *Discrete Dynamics in Nature and Society*, Vol. 2017, 9741824, pp. 1-7.

Zager, L.A. and Verghese, G.C. (2008), "Graph similarity scoring and matching", *Applied Mathematics Letters*, Vol. 21 No. 1, pp. 86-94.

Zhang, X., Liu, X., Li, X. and Pan, D. (2017), "MMKG: an approach to generate metallic materials knowledge graph based on DBpedia and Wikipedia", *Computer Physics Communications*, Vol. 211, pp. 98-112.

Zhang, Y., Chen, H., Lu, J. and Zhang, G. (2017), "Detecting and predicting the topic change of knowledge-based systems: a topic-based bibliometric analysis from 1991 to 2016", *Knowledge-Based Systems*, Vol. 133, pp. 255-268.

Corresponding author
Shwe Sin Phyo can be contacted at: shwesinphyo@ucsm.edu.mm