Received 27 October 2016
Revised 8 June 2017
Accepted 1 August 2017

Considering social information in constructing research topic maps

Hei Chia Wang, Yu Hung Chiang and Yen Tzu Huang
Department of Industrial and Information Management and
Institute of Information Management, National Cheng Kung University,
Tainan, Taiwan

Abstract

Purpose – In academic work, it is important to identify a specific domain of research. Many researchers may look to conference issues to determine interesting or new topics. Furthermore, conference issues can help researchers identify current research trends in their field and learn about cutting-edge developments in their area of specialization. However, so much conference information is published online that it can be difficult to navigate and analyze in a meaningful or productive way. Hence, the use of knowledge management (KM) could be a way to resolve these issues. In KM, ontology is widely adopted, but most ontology construction methods do not consider social information between target users. Therefore, this study aims to propose a novel method of constructing research topic maps using an open directory project (ODP) and social information.

Design/methodology/approach – The approach is to incorporate conference information (i.e. title, keywords and abstract) as sources and to consider the ways in which social information automatically produces research topic maps. The methodology can be divided into four modules: data collection, element extraction, social information analysis and visualization. The data collection module collects the required conference data from the internet and performs pre-processing. Then, the element extraction module extracts topics, associations and other basic elements of topic maps while considering social information. Finally, the results will be shown in the visualization module for researchers to browse and search.

Findings – The results of this study propose three main findings. First, creating topic maps with the ODP category information can help capture a richer set of classification associations. Second, social information should be considered when constructing topic maps. This study includes the relationship among different authors and topics to support information in social networks. By considering social information, such as co-authorship/collaborator, this method helps researchers find research topics that are unfamiliar but interesting or potential cooperative opportunities in the future. Third, this study presents topic maps that show a clear and simple pathway in interested domain knowledge.

Research limitations implications – First, this study analyzes and collects conference information, including the titles, keywords and abstracts of conference papers, so the data set must include all of the abovementioned information. Second, social information only analyzes co-authorship associations (collabship associations); other social information could be extracted in the future study. Third, this study only analyzes the associations between topics. The intensity of associations is not discussed in the study.

Originality/value – The study will have a great impact on learned societies because it bridges the gap between theory and practice. The study is useful for researchers who want to know which conferences are related to their research. Moreover, social networks can help researchers expand and diversify their research.

Keywords Text mining, Social information, Open directory projects, Topic maps

Paper type Research paper

The Electronic Library
Vol. 36 No. 2, 2018
pp. 220-236
© Emerald Publishing Limited
0264-0473
DOI 10.1108/EL-10-2016-0230

The research was based on work supported by the Taiwan Ministry of Science and Technology under Grant N. MOST 103-2410-H-006-055-MY3.

Introduction

In the realm of information retrieval, research capability has come to be seen as the source of knowledge (Numprasertchai and Igel, 2005). Researchers are expected to know about new research trends and to cooperate with other researchers across disciplines (Palmer, 2013, Tomaszewski and MacDonald, 2009). With the large number of research articles that have been converted to electronic format, many researchers try to identify interesting research topics by reviewing online resources. Finding a proper research domain is an important issue for these researchers. During the search process, conference papers could be one of the greatest potential resources because most conference program chairs are likely to consider new trends for conference topics. Unlike journals, which have stricter editing guidelines and a longer publishing period, the conference publishing period is normally shorter and closer to the current focus. In conferences, researchers can realize current research trends in their field and learn about cutting-edge developments in their specializations. Moreover, conferences play an important role in scholarly communication because they provide scientists and scholars with an opportunity to present and discuss the preliminary results of their research and to improve their personal social networks (González-Albo and Bordons, 2011).

Although many conference resources are available on the Internet on sites, such as DBworld and ConfSearch, most of them must still be searched via keywords. There are no detailed or unified classifications for conferences or the possible relationships between classifications. For example, currently, the more popular conference resources, such as AllConferences.com (www.allconferences.com) and conference alerts (www.conferencealerts.com), offer only manual conference knowledge classifications or no classification at all. They are not designed to include domain knowledge or possible relationships between classifications. This condition leads to incomplete search results. Researchers spend lots of time in finding the information they need. Due to the time limitation, some useful information may not be found. Therefore, a conference information analysis platform could be designed to help researchers find suitable research domain-related topics.

To this end, knowledge management (KM) could be a viable resource (Biswas *et al.*, 2014; Hamasaki *et al.*, 2007). Conference content, such as keywords, articles and websites, could be properly analysed and classified by the KM method. Constructing a conceptualized architecture would allow researchers to extend topics of interest and improve the accuracy of their search results. In recent years, topic maps, which have been widely used in the construction of conceptual architecture, have received growing attention (Santoso *et al.*, 2011). They can transform implicit knowledge into explicit knowledge, providing benefits to many applications – for example, information retrieval systems (Lee *et al.*, 2007; Xia *et al.*, 2016) – and Web search engines (Al-Rajebah and Al-Khalifa, 2010; Chiu and Pan, 2014). There have been many studies concerning topic map construction (Du *et al.*, 2009; Santoso *et al.*, 2011; Yao *et al.*, 2013). However, existing topic map construction methods do not consider social information.

During the past decade, social network sites have become one of the most popular access points. In social networks, access records can be used to determine relationships among social members (Hu and Racherla, 2008; Vigneshwari and Aramudhan, 2015). Hamasaki *et al.* (2007) suggested that considering social information, such as who collaborates and/or cooperates with whom, might improve the extracted conceptual architecture. Therefore, the purpose of this study is to incorporate conference information (i.e. title, keywords and abstract) as sources and to consider the ways in which social information automatically produces research topic maps. By providing the relationships between topics, this study hopes to offer researchers more complete references when accessing conference information.

Literature review

Social information and knowledge management

Many researchers consider social networks to be an important factor in understanding the knowledge creation process (Cugmas *et al.*, 2016). Social networks define the correlation between community members, and this correlation could directly affect cooperation and exchange of knowledge among members (Hu and Racherla, 2008). Social network analysis (SNA), which can be described by graphs, is based on relationships between community members (Liu *et al.*, 2005). It is used to analyse the characteristics and patterns between specific domain members (Hu and Racherla, 2008).

In SNA, co-authorship is regarded as a kind of social information. Co-authorship creates a social network, the study of which allows us to understand the structure of scientific collaborations and some of the characteristics of a particular discipline (Acedo *et al.*, 2006). Many researchers have opined that co-authored papers are of better quality than single-authored papers. Hudson (1996) describes *co-authorship* as the participation of two or more authors in the production of a study, which leads to a scientific output of a greater quality or quantity than could be achieved by a single individual. Achieving scientific output of increased quality and quantity is one of the reasons for the rising trend in co-authorship (Acedo *et al.*, 2006). This argument stems from the fact that as a result of the rise in the complexity of all disciplines, it becomes necessary to combine the skills of two or more researchers who are specialists in their various fields to improve output quality.

Co-authorship networks have been an important class of social networks and have been used extensively to determine the structure of scientific collaborations and the status of individual researchers (Liu *et al.*, 2005). Liu *et al.* (2005) also noted that co-authorship has a stronger social relationship than citing another author. Citations can occur without the authors knowing each other and can span across time. Co-authorship implies a temporal and collegial relationship that places it more squarely in the realm of social network analysis and receives more attention in SNA.

Ontology and topic maps

The word *ontology* is derived from the field of philosophy, with the important specification that it has great potential to help in the organizing and managing of knowledge (Santoso *et al.*, 2011). Ontology can be used to present knowledge hidden within a mountain of data; it can integrate domain knowledge and demonstrate its consistency (Jiang *et al.*, 2005; Xia *et al.*, 2016). In recent years, ontology has played an increasingly important role in textual analysis and information exchange between different areas (Zhai and Massung, 2016).

Topic maps are used as a formal syntax for the representation and implementation of ontologies; they can be used to find information and are suitable for non-scientific areas (Yi, 2008; Zhai and Massung, 2016). Topic maps are a technology for the encoding of knowledge and connecting this encoded knowledge to relevant information resources (Kim *et al.*, 2007). Currently, topic maps have been developed into XML topic maps (XTM), and studies about topic maps continue to develop rapidly.

Topic maps model

Topic maps are composed of three elements: topics, associations and occurrences (TAO). *Topics* are subjects in knowledge domains. *Associations* are relationships between subjects, while *occurrences* connect subjects and related information resources (Kim *et al.*, 2007). A detailed description of TAO is as follows (Pepper, 2010):

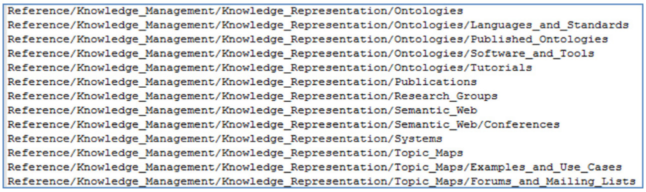
- *Topics*. In general, a topic can be any "thing" – a person, an entity, a concept – regardless of whether it exists or has any other specific characteristics.
- *Associations*. A topic association asserts a relationship between two or more topics. For example, Puccini "was born in" Lucca or Lucca "is in" Italy.
- *Occurrences*. A topic may be linked to one or more information resources that are deemed to be relevant to the topic in some way. Such resources are called occurrences of the topic.

Classification resource – open directory project. Many studies use a semantic knowledge base, such as WordNet, to identify meaning, related terms and other information (Oliva *et al.*, 2011). However, the information is usually too general to extract specific domain associations (Sánchez and Moreno, 2008). For example, researchers cannot find "semantic web" or "ontology" in WordNet. However, this study uses a corpus in the field of computer science; thus, general semantic knowledge is less suitable. This study relies on a background of specific domain knowledge to help to extract taxonomy associations, and this study, therefore, selects Open Directory Project (ODP) to extract professional domain associations.

ODP is an open category that is manually constructed and maintained. It has 16 main categories and more than one million subcategories (Jiang and Tan, 2009). Figure 1 shows several ODP categories, each column representing a particular category. For example, the first line of categories presents ontologies, language and standards, published ontologies and software and tools. According to recent news, all the data in ODP have become a static archive. ODP ceased allowing further editing in March 2017 because of its owner, American OnLine (AOL), no longer wishing to support the project.

Evaluation of topic maps

Evaluating ontologies is an important issue (Brank *et al.*, 2005; Wu *et al.*, 2015). Errors and omissions may cause applications to fail to fulfil their information exchange potential, while a good ontology can render the ontology reusable and establish cooperation between a domain and its applications (Venkatesh *et al.*, 2007). This study refers to *accuracy* (ontology is correct and represents the real world), *clarity* (ontology is correct and represents the real world) and *completeness* (whether the ontology covers a given field of interest appropriately; whether it contains all of the relevant concepts and vocabularies, also known as *richness*) (Vrandečić, 2009), furthermore adding *ease of use* (whether it is easy to use), *efficiency* (whether it can find the required information quickly) and *satisfaction* (whether researchers are satisfied with the system), giving a total of six criteria for evaluation.



```
Reference/Knowledge_Management/Knowledge_Representation/Ontologies
Reference/Knowledge_Management/Knowledge_Representation/Ontologies/Languages_and_Standards
Reference/Knowledge_Management/Knowledge_Representation/Ontologies/Published_Ontologies
Reference/Knowledge_Management/Knowledge_Representation/Ontologies/Software_and_Tools
Reference/Knowledge_Management/Knowledge_Representation/Ontologies/Tutorials
Reference/Knowledge_Management/Knowledge_Representation/Publications
Reference/Knowledge_Management/Knowledge_Representation/Research_Groups
Reference/Knowledge_Management/Knowledge_Representation/Semantic_Web
Reference/Knowledge_Management/Knowledge_Representation/Semantic_Web/Conferences
Reference/Knowledge_Management/Knowledge_Representation/Systems
Reference/Knowledge_Management/Knowledge_Representation/Topic_Maps
Reference/Knowledge_Management/Knowledge_Representation/Topic_Maps/Examples_and_Use_Cases
Reference/Knowledge_Management/Knowledge_Representation/Topic_Maps/Forums_and_Mailing_Lists
```

Figure 1.
Portion of ODP categories

Methodology

This section describes the proposed method of constructing topic maps while considering social networks and knowledge. The research architecture, as shown in Figure 2, can be divided into four modules: data collection, element extraction, social information analysis and visualization. The data collection module collects the required conference data from the internet and performs pre-processing, such as part-of-speech tagging, to enable the follow-up module to perform an analysis. Then, the element extraction module extracts topics, associations and other basic elements of topic maps while considering social information. Finally, the results can be visualized in the visualization module for researchers to browse and search. The following describes the detailed process and content of the modules, as well as the research theory and methods.

Data collection module

The data collection module is divided into two parts: data collection and data processing. The source comes from the Association for Computing Machinery (ACM) Digital Library proceedings on the Internet. The corpus of this study collects articles in conferences (Ci), where Ci represents conference i and Docij represents the jth article in conference i, Docij ∈ Ci. Shah *et al.* (2003) compared the full text to the abstract and found that the abstract presents important information for understanding the study and that it is sufficient for presenting the content of the study. Thus, the work presented herein will analyse the title, keywords and abstract. Furthermore, this module also conducts social network processing and collects information on conference authors and co-authorship for the remaining modules. Every article is composed of a title (t), keyword (k), abstract (a) and scholar (s), and it is represented as Docij = {tij, kij, aij, sij}. Here, sij = {dijh}, dijh represents the hth author of the jth article in conference i.

The system then uses natural language processing to perform pre-processing, including part-of-speech (POS) tagging, stop words and stemming. First, it performs POS tagging on the titles and abstracts using a natural language analysis tool, the Stanford Parser, to retrieve nouns or noun phrases. For example, the result of POS tagging is NN (Noun,

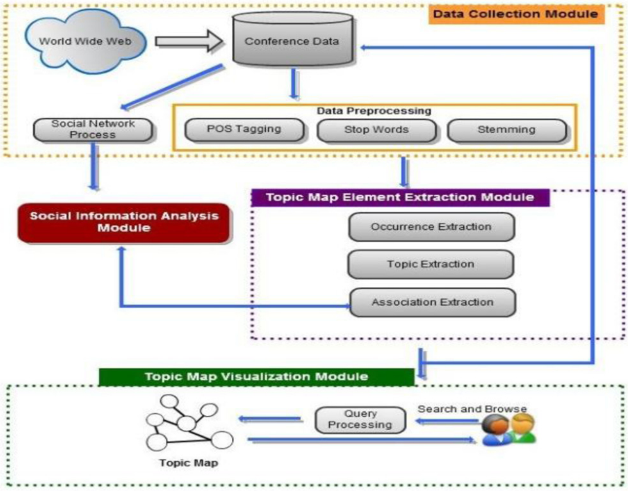


Figure 2.
Research architecture

singular or mass) or NNS (Noun, plural). After this step, the study can obtain title noun sets and abstract noun sets. Then, the study uses a stop word corpus to remove stop words, such as *a*, *the* and *is*. The third step is stemming, using Porter's stemmer to stem previously captured terms. This step avoids mistakes that have been introduced by plural and singular words that have the same meaning but are recognized as different words. After stemming, the title noun sets, (tNounij), tNounij = {t_Wijm | where t_Wijm ∊ noun and t_Wijm ∊ tij}; abstract noun sets (aNounij), aNounij = {a_Wijn | where a_Wijn ∊ noun and a_Wijn ∊ aij}; and keyword sets (kij), kij = {k_Wij1, k_Wij2,. . ., k_Wijp}, where t_Wijm (a_Wijn) is the () th noun in the title (abstract) and p is the number of keywords, are obtained.

Topic maps element extraction module

After data pre-processing, the topic maps element extraction module is used. This module extracts feature word sets from conference articles and finds connections between topics. There are three separate parts of this module: topic extraction, association extraction and occurrence extraction. The following section describes these parts in detail.

Part I. Topic extraction. The topics that were extracted in this study can be divided into four categories: conference topics, conference names, ODP classes and authors of conferences. Except for the conference topics, other topic categories can be derived from a data collection module or ODP classification source. The method of extracting conference topics uses the concept of term frequency (TF) to find the feature set of every conference. The steps of topic extraction were as follows:

- *TF calculation.* This study uses the TF formula and references the method of (Du *et al.*, 2009), giving different classes different weights and giving different class weights according to where the terms appear in the article (i.e. title, keyword and abstract), in which cw_1, cw_2 and cw_3 represent the weights of the title, keyword and abstract, respectively, in sequence. TF is calculated according to the classes and summarized. The numerator is the frequency of the term that appears in the class, while the denominator is the total number of terms that appear in the class. Summarizing each weighted frequency of three classes in each of the articles in the conference, this study obtains the representativeness of the term – that is, its total weighted frequency (wf) – as in Formula (1), where cwb is the weight of class b:

$$wf = TermFreq_b gClassWeight = \sum_{b=1}^{3}(tf_b \times cw_b) \tag{1}$$

- *Normalization.* The variable wf is normalized according to Formula (2). Here, wf'_g is the importance of term g in the conference, and wnum is the total number of terms in all conference documents:

$$wf'_g = \frac{wf_g - \min(wf_h)}{\max(wf_h) - \min(wf_h)}, h = 1 \ldots wnum \tag{2}$$

- *Filter terms*. After calculating all wf'_g terms, the study referred to the ACM Digital Library website, which defined approximately 15 topics. Therefore, the study captured the top 20 topics.

Part II. Association extraction. An association is a relationship between topics. Topic maps can describe the association of terms that belong to the same hierarchy and/or different hierarchies (Yi, 2008). Sánchez and Moreno (2008) believed that relationships between concepts in ontology can be taxonomically-related or non-taxonomically-related. Taxonomically related means an "is-a" relationship, such as a local area network (LAN) "is a" type of network. Non-taxonomically-related relationships refer to naming; for example, Pepper "is the author of" topic maps. Sánchez and Moreno (2008) also noted that studies of ontology to date have often focused on taxonomical relations and ignored non-taxonomical relations. Thus, the associations that are extracted by this study can be divided into taxonomical associations and non-taxonomical associations:

- Taxonomical association. The relation between topics belongs to the same hierarchy, such as related terms, hypernyms and hyponyms.
- Non-taxonomical association. The relations between topics belong to different hierarchies, such as authorship being between the author and topic and co-authorship being between authors.

Non-taxonomic association will be introduced in the social information analysis module; here, this study introduces taxonomic association. This study extracts seven different taxonomic associations in total, where RT, BT and NT associations refer to the definitions of Slawsky (2007); this study extracts RT (is related to), BT (broader term), NT (narrower term), ST (sibling term), ONT, OBT and OST, which can be described as follows:

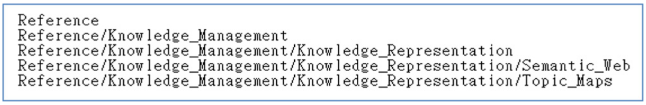
- RT: Conference topic is related to the ODP class;
- BT: ODP class is the upper level (hypernym) of another ODP class;
- NT: ODP class is the lower level (hyponym) of another ODP class;
- ST: The association of ODP classes has the same parent class;
- ONT: The association between conference topics generated by association NT between ODP classes;
- OBT: The association between conference topics generated by association BT between ODP classes; and
- OST: The association between conference topics generated by association ST between ODP classes.

Because a specific domain knowledge background is required to assist in the extraction of taxonomical association, this study uses ODP to extract the hierarchical relations among topics. Figure 3 shows part of the ODP class. Each column is a specific class. The level of the terms in ODP is determined by the number of occurrences of the symbol "/" plus 1. For example, Level (Reference) = (0 + 1) = 1 and Level (Topic_Maps) = (3 + 1) = 4.

The steps of the association extraction process are as follows:

- *RT association.* The similarity of the conference topic and ODP class is calculated based on the Levenshtein distance (LD). If the value is larger than the threshold, the RT association is defined between two topics, Tx and Ty. The similarity formula is

Figure 3.
A portion of the ODP class



```
Reference
Reference/Knowledge_Management
Reference/Knowledge_Management/Knowledge_Representation
Reference/Knowledge_Management/Knowledge_Representation/Semantic_Web
Reference/Knowledge_Management/Knowledge_Representation/Topic_Maps
```

given in Formula (3). LD (Tx, Ty) is the edit distance between two topic strings, in other words, the minimum numbers of insertions, deletions and substitutions that are involved in converting Ty to Tx. Here, max (|Tx|,|Ty|) is the maximum length of two strings:

$$RT_{sim}(T_x, T_y) = \frac{\max(|T_x|, |T_y|) - \mathrm{LD}(T_x, T_y)}{\max(|T_x|, |T_y|)} \tag{3}$$

- *The topic list is generated.* Every ODP class is a topic. Every ODP topic is linked: ODP_Tx, which is related to the conference topic, to become the topic list. Here, onum is the total number of related ODP classes. Topic List = {ODP_Tx|where 1 x onum}.
- *Topic pairs are generated.* Topic pairs (ODP_Tx, ODP_Ty) exist between any two topics in the topic list.
- *Associations are extracted.* Each topic pair is compared. If the following situation occurs, the BT, NT and ST associations are extracted between the two topics:
  - When the topics in the topic pair appear in the same column of the ODP hierarchy and Level (ODP_Tx) > Level (ODP_Ty), ODP_Tx is the BT of ODP_Ty;
  - When the topics in the topic pair appear in the same column of the ODP hierarchy and Level (ODP_Tx) < Level (ODP_Ty), ODP_Tx is the NT of ODP_Ty; and
  - When Level (ODP_Tx) = Level (ODP_Ty) and the topics have the same parent class, ODP_Tx is the ST of ODP_Ty.
- The BT and NT associations are symmetric relations; specifically, if ODP_Tx is the BT of ODP_Ty, then ODP_Ty is the NT of ODP_Tx and vice versa.
- *The ST association is calculated.* When the extracted association is ST, the degree of ST association must be calculated because even with the same parent class, it could be the same first parent class or the same parent class after several levels, which indicates a different similarity. The calculation method of this study was developed by referring to and modifying Lee *et al.* (2007)'s concept of calculation similarity between terms using the path length. The path length is defined as follows:
  - Path_length = length of the path between two topics in the ODP structure; and
  - The distances of the topics to the same parent class are summed, and then, using normalization, this study can obtain the ST association similarity degree (STsim), as depicted in Formula (4). MaxPL is the maximum path length in the ODP hierarchy, and MinPL is the minimum path length in the ODP hierarchy:

$$ST_{sim} = \frac{Max_{PL} - path\_length}{Max_{PL} - Min_{PL}} \tag{4}$$

- *The conference topic association is extracted.* After extracting the association RT between the ODP class and conference topics and the BT, NT and ST associations between the ODP classes, this study can then extract new OBT, ONT and OST associations between two topics that have an RT association with the ODP class, and the new association is generated by an association among ODP classes.

Part III. Occurrence extraction. A topic can link to one or more information resources that are related to topics; these resources are called *occurrences of the topic*. This study extracts conference information from the Internet; if the topic is a conference, the occurrence of the topic (T_OCCi) is composed of websites and places. CWebi is the website of conference i and CPlacei is the place of conference i, as follows:

$$\mathrm{T\_OCCi} = \{\mathrm{CWebi},\ \mathrm{CPlacei}\}$$

Social information analysis module

The data collection module will determine the authors of the articles and those authors' published articles. The social information analysis module will offer better results by adding social information or offering research direction or inspiration when researchers search. As interdisciplinary research collaboration is increasing, this study suggests that the topic author and co-authorship of interest share a development space, such as a common combination of biological aspects and information. Therefore, the module analyses the social information that is related to topic association, finds related topics by considering social information and extracts non-taxonomic associations.

The non-taxonomic associations that this study extracts are the "is the author of" associations between authors and topics and the "collabship" associations among the topics generated by co-authorship. In academics, co-authorship may be the most important type of link between researchers (Hwang *et al.*, 2010). According to the data collection module, this study can determine the authors of articles and their co-authors, and through the topic extraction step, this study can determine the topics that represent each article. Thus, two associations can be extracted:

(1) *Is the author of association.* When author (sr) publishes an article, he/she will have an "is the author of" association with the topics of the article; here, sr is the author of Tijx, and Tijx is the x-th topic of the jth article in conference i.

(2) *Collabship association.* Two different topics from the same article have a collabship association because of co-authorship between the authors. Tx has collabship with Ty, and Tx and Ty come from the same article, Tx ≠ Ty.

Topic map visualization module

Using the methods of the previous phases, this study can obtain the elements (topic, association and occurrence) that are needed to construct topic maps and visualize them in this module via tools. Researchers can search and browse topics and conference information. Eventually, they can find the structure and associations of the topic maps, as depicted in Figure 4. For example, KM is the BT of ontology, or Rose ER is the author of information retrieval, or ontology and biology have a collabship association. Additionally, the topics will link to the relevant occurrences, such as the website and location of the conference.

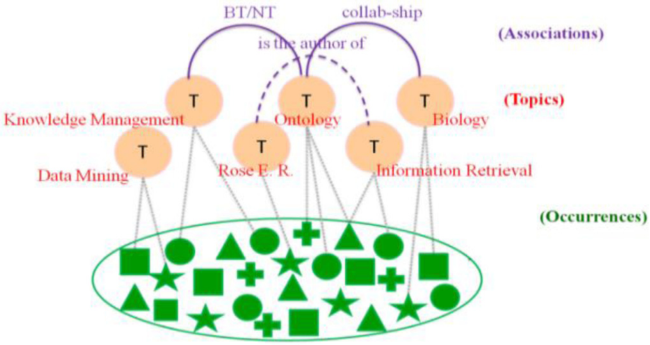
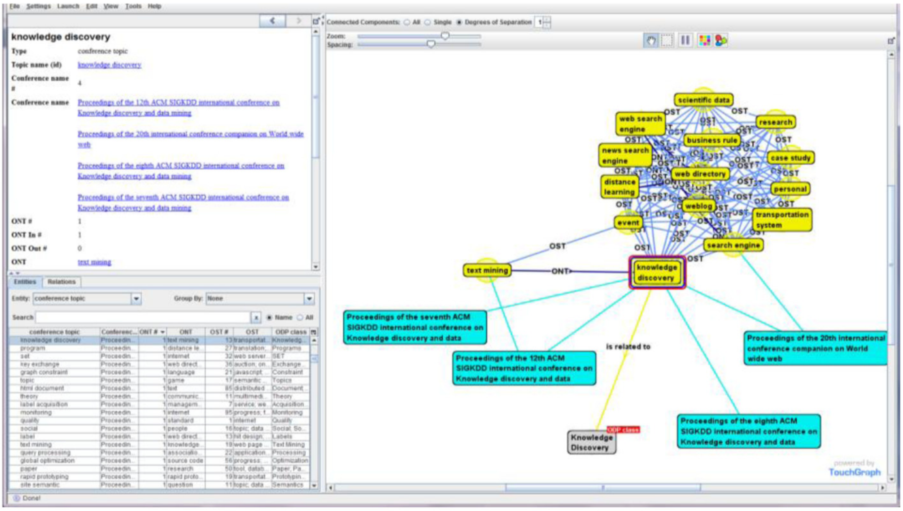


Figure 4.
The elements of topic maps

Implementation and evaluation

The implementation discussed in this study is built and operated in the operating system of Ubuntu 10.04 and Win 7 64-bit, using Perl 5.10.1 as the programming language and MySQL 5.1.6 as the database. The hierarchy classification source used here, called ODP, contains many classes that differ from one another; thus, this study selects 16 main categories and their subcategories, choosing those that are related to information science. The categories involve business, computers, reference and science.

This study uses the proposed method to extract the elements of topic maps and visualize them, as shown in Figure 5. Each node is a topic that could represent the conference name, conference topic, ODP class, or any of the information extracted in this study. Different categories will be shown in different colours on the graph. The top left block is the description, which contains the topic description and other relevant information. The bottom left block is represented in table form. Researchers can search in the block according to different topic types, such as the conference name or topic. Researchers can also search for


File Settings Launch Edit View Tools Help
Connected Components: ○ All ○ Single ● Degrees of Separation 1
Zoom:
Spacing:
Entities
Relations
scientific data
web search engine
research
business rule
news search engine
case study
web directory
distance learning
personal
weblog
transportation system
event
search engine
OST
text mining
ONT
knowledge discovery
is related to
ODP class
Knowledge Discovery
Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data
Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data
Proceedings of the 20th international conference companion on World wide web
Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data
powered by TouchGraph

knowledge discovery

Type	conference topic
Topic name (id)	knowledge discovery
Conference name #	4
Conference name	Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining Proceedings of the 20th international conference companion on World wide web Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining
ONT #	1
ONT In #	1
ONT Out #	0
ONT	text mining

Entity: conference topic Group By: None

Search ● Name ○ All

conference topic	Conferenc...	ONT #	ONT	OST #	OST	ODP class
knowledge discovery	Proceedin...	1	text mining	13	transportat...	Knowledg...
program	Proceedin...	1	distance le...	27	translation;...	Programs
set	Proceedin...	1	internet	32	web server...	SET
key exchange	Proceedin...	1	web direct...	36	auction; on...	Exchange...
graph constraint	Proceedin...	1	language	21	javascript;...	Constraint
topic	Proceedin...	1	game	17	semantic...	Topics
html document	Proceedin...	1	text	85	distributed...	Document...
theory	Proceedin...	1	communic...	11	multimedi...	Theory
label acquisition	Proceedin...	1	managem...	7	service; we...	Acquisition...
monitoring	Proceedin...	1	internet	95	progress; f...	Monitoring
quality	Proceedin...	1	standard	1	internet	Quality
social	Proceedin...	1	people	16	topic; data...	Social; So...
label	Proceedin...	1	web direct...	13	hit design;...	Labels
text mining	Proceedin...	1	knowledge...	19	web page...	Text Mining
query processing	Proceedin...	1	associatio...	22	application...	Processing
global optimization	Proceedin...	1	source code	56	progress;...	Optimization
paper	Proceedin...	1	research	50	tool; datab...	Paper; Pa...
rapid prototyping	Proceedin...	1	rapid proto...	19	transportat...	Prototypin...
site semantic	Proceedin...	1	question	11	topic; data...	Semantics

Done!

Figure 5.
Example of the system interface

any association, such as RT or "is the author of". Topic maps are in the right block and can be double-clicked. The tools above topic maps can zoom in, zoom out and change the spacing between topics.

Source and experiment methods

This study extracted conference information in the computer science field from the ACM Digital Library Proceedings. Because ACM contains many documents, this study refers to the computer science conference ranking summary and selects five fields and six conferences from 2000 to 2011, which yielded 7,760 articles in total.

The evaluation of this study can be divided into subjective evaluation and objective evaluation. Subjective evaluation uses questionnaires for research. This study refers to the Vrandečić (2009) ontology quality standard in Table I, with a design task questionnaire for experts and researchers to evaluate the effectiveness of the study. An objective evaluation must use the concept of coverage; coverage can refer to the coverage compared to Wikipedia, as stated by Ponzetto and Strube (2011). The present work defines the coverage as the ratio of conference topics that appear in ACM topic fields, as shown in Formula (5). |EACM| is the number of terms that appear in the ACM topic fields, and |ACM| is the number of terms in the ACM topic fields:

$$Coverage_{topic} = \frac{|E_{ACM}|}{|ACM|} \times 100\% \tag{5}$$

The weight analysis in this study referred to the setting of the title, keyword and abstract of Wang *et al.* (2009). The weights are shown in Table II. According to the weights in the five different groups to calculate the topic maps coverage in the ACM topics field, the present work will choose one group of weights to construct topic maps and the task questionnaire evaluation.

Standards	Description
Ease of use	Whether system is easy to use
Efficiency	Whether can find the required information quickly
Completeness	Whether information covers interested field appropriately and whether it contains all the relevant concepts and vocabularies
Accuracy	Whether system can offer correct information and whether it contains irrelevant information
Clarity	Whether the information offered by the system is easy to understand
Satisfaction	Whether researchers are satisfied with system use

Table I.
Subjective evaluation standards

	cw_1 title weight	cw_2 keyword weight	cw_3 abstract weight
Group 1	2	4	1
Group 2	2	2	1
Group 3	1	2	1
Group 4	1	1	1
Group 5	0.32	0.35	0.33

Table II.
The parameters of weight analysis

This study, while referring to Venkatesh *et al.* (2007), compared the search efficiency of topic maps and generated three tasks. The descriptions of each task were as follows:

- Task 1: The respondents are researchers who want to submit to a conference and want to know what the conferences are about with respect to a particular topic;
- Task 2: The respondents are researchers who want to submit to a conference and want to know what topics the conference has published on; and
- Task 3: The respondents are researchers who want to submit to a conference and have a specialty area; they want to know about the related topics to find a possibility for cooperation among topics or authors.

The respondents searched and answered on both the ACM website and topic maps, and they performed an evaluation regarding the utility of each association. The questionnaire was divided into five options, including strongly agree (5 points), no comment (3 points) and strongly disagree (1 point).

A 5-10 min introduction was provided to explain the research and system for the respondents before they answered the questionnaires. To avoid order effects, half of the questionnaires asked about ACM first, and the other half had asked about topic maps (TM) first. In addition, all of the operating actions and processes were recorded with the agreement of the respondents. To avoid bias introduced by considering only a single topic, the questionnaire offered randomly chosen topics for the respondents to choose from.

This study contained a total of three experiments. The first concerned comparative thematic areas of coverage rate to obtain the best weight implemented as a topic map. The second and the third related to the analysis of questionnaire data for the ACM and topic maps, respectively, as well as the benefits and an associated performance assessment of the topic maps. A total of 52 questionnaires were filled out by National Cheng Kung University students who studied in the information engineering or information management departments or were contributors or participants in the research conferences. Of these, 70 per cent of the respondents were male and 30 per cent were female; the respondents were between the ages of 22 and 27 years old, with an average age of 23.9. In this group, 76 per cent of the respondents took part in or contributed to the experience of the conferences, and 56 per cent of the respondents viewed the ACM Digital Library website. The administration of each questionnaire included commentary, the actual operation and completion of the questionnaire. The participants spent an average of approximately 3 min on the study. Cronbach's α reliability index was used for the questionnaire; the part of the questionnaire concerning the ACM website had a Cronbach's α value of 0.813, and the part of the questionnaire concerning the topic maps system had a Cronbach's α of 0.772. These values are in compliance with the requirement that the Cronbach value should be greater than 0.7 to satisfy the internal consistency test. The following sections describe the three experiments.

The analysis of weight and coverage rate

The purpose of this experiment is to understand whether the method of the study can capture the appropriate topic to cover six conferences (for all of the years) offered by the ACM and obtain the best weight to implement topic maps. This research computed the weights of five groups in Table II to extract the topic of the conferences by full-word matching combined with partial-word matching, covering the ACM subject areas, as shown in Figure 6. The figure shows, with full compliance in this case, a coverage rate up to 78.03 per cent, with the first group and third group performing the best.

If the topic is more than one word in the ACM subject areas, the coverage rate is increased to more than 90 per cent of the individual words, compared with a good percentage for the first and third groups, the coverage rate remained higher than in the other groups. This method draws from Wang *et al.* (2009), who used the weights of the title, keywords and abstracts for their experimental results. Their results showed no significant difference in the importance of the keywords and title but also showed that they are more important than the summary. This study took the first sets of 2, 4 and 1 to set the title, keywords and abstracts, respectively, of the three weights that were used to implement the topic maps.

Analysis of subjective evaluation

The purpose of this experiment is to understand the performance of topic maps with respect to the assessment indicators; the analysis of questionnaire data; and the mean value, standard deviation and range, as shown in Table III. In terms of the three similar situations, Task 2 (according to the conference to find published topics) corresponds with the identification of better performance when the published subject is considered.

Furthermore, this study is intended to provide further insight as to whether there are significant differences between the TM and the ACM in all situations. Thus, the statistical *t*-test was used, with μ 1 for the TM of the population mean, μ the population mean for the ACM, H 0: μ 1 = μ 2; H 1: 1 $\neq$ μ 2 and a confidence interval of 0.95. The results show that the accuracies between Task 1 and Task 2 and other assessment projects (ease of use, efficiency, integrity, clarity and satisfaction in all situations) have significant differences.

In Tasks 1 and 2, the accuracy is not significant; this result can be explained by the fact that the TM and ACM have the same accuracy. As all TM data are extracted from ACM websites, they have a certain degree of correlation. This study expects to propose a different information structure that helps researchers get more useful and rapid search results. In the three tasks, especially those that address situational performance, the difference in the most significant reason for Task 3 provides information on how the research community seeks thematic cooperation or the possibility of cooperation with other authors. Thus, from the results of this study, it can be inferred that the researchers of the TM can more easily and quickly find the information that is needed and that the information presented by the TM is easy to understand.

Performance of all associations

The purpose of this study is to understand if topic maps can help researchers implement the performance of all associations. The analysis of the questionnaire data, including the

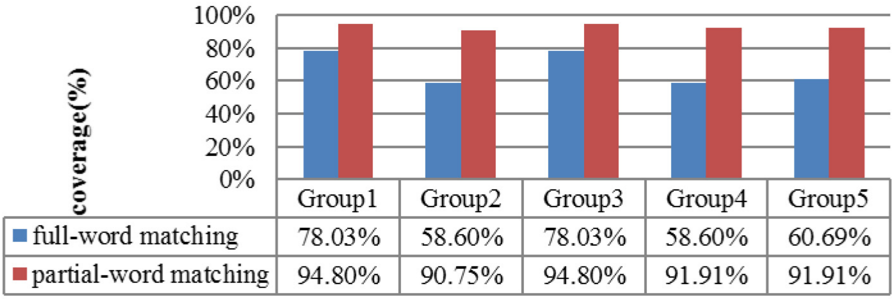


Figure 6.
The results of coverage rate with different weights

Task measure	T1 ACM	 TM	T2 ACM	 TM	T3 ACM	 TM
Ease of use						
Mean	3.44	_4.28_	3.22	_4.36*_	3.3	_4.16_
SD	0.93	0.67	1.09	0.6	1.01	0.77
Efficiency						
Mean	3.2	_4.24_	2.98	_4.42*_	2.6	_4.34_
SD	0.99	0.74	1.04	0.7	1.21	0.72
Completeness						
Mean	3.56	_4.14*_	3.62	_4.02_	2.96	_3.8_
SD	1.28	0.85	0.95	0.82	1.08	0.95
Accuracy						
Mean	4.02	_4.1*_	3.92	_4_	3.5	_4.1*_
SD	0.47	0.58	0.63	0.53	0.79	0.54
Clarity						
Mean	3.14	_4.18_	3.38	_4.12_	3.24	_4.22*_
SD	1.13	0.94	1.03	0.85	1.06	0.89
Satisfaction						
Mean	3.42	_4.1_	3.04	_4.14*_	2.9	_4.12_
SD	0.84	0.61	1.03	0.67	0.96	0.48

Notes: Italic text indicates a higher average, and the table shows that the TM performance is better in most cases; an asterisk * notes the overall best performance as determined by the evaluation indicators

Table III. Data from the second experiment

distribution, mean and standard deviation, is shown in Table IV. In the associated ONT and OBT for the reverse relationship, this study measures only the results that can be calculated, with the questionnaires and analysis showing only the results of the ONT. The denominator represents the 50 questionnaires, and the numerator option is selected as the number of figures in the table. From Table IV, the associated topic maps capture the average performance of all four points. The majority of the respondents agreed that these associations were helpful to them. Among them, RT, collabship and "is the author of" performed better. It can be inferred that the joined RT is a link to the conference topic and the ODP categories that are associated with the ODP category information in the conferences information search assist researchers in their understanding of conference information. The other is an author who has a collabship association and is established in consideration of the community information; hence, this study infers the consideration of social information to help researchers to understand research information.

Score association	5	4	3	2	1	Mean	SD
ONT	15/50	30/50	5/50	0/50	0/50	4.22	0.58
OST	10/50	35/50	5/50	0/50	0/50	4.12	0.53
RT	25/50	24/50	1/50	0/50	0/50	4.51	0.51
Is the author of	17/50	25/50	6/50	0/50	0/50	4.24	0.63
Collabship	23/50	20/50	7/50	0/50	0/50	4.35	0.70

Table IV. The results of the performance of all associations

Conclusion and future directions

This study proposes a method of constructing topic maps using the ODP and social information to help researchers obtain more useful search results from conference information. The results of this study propose three main findings. First, creating topic maps with ODP category information can help capture a richer set of classification associations. Second, previous studies have not considered social information between authors. This study includes the writing relationship among different authors and topics to support information in social networks. By considering social information, such as co-authorship/collaborator, this method helps researchers find research topics that are unfamiliar but interesting or potential cooperative opportunities in the future. This useful social information offers more information than general conference issues. Third, this study presents topic maps that show a clear and simple pathway to domain knowledge.

According to Experiment 1, the first group has a better coverage rate than other groups (78.03 per cent and 94.8 per cent, respectively) (2, 4 and 1). This study was given weight in the title, keywords and abstract, and according to the result of questionnaire analysis in Experiments 2 and 3, TM and ACM sites have significant differences in their performance on the assessment indicators. The cause of these significant differences is based on knowledge structure topic maps that can be used more easily and quickly to find needed information. Moreover, the mean performance score of all associations is greater than four points. This result shows that the majority of respondents agreed that these associations are helpful to them because it is difficult to obtain this information from general conference issues.

Regarding future directions, there is much social information conveyed between authorships, but this study only extracts "collabship" associations (one kind of social information); other social information, such as times cited, could be considered in further studies. Topic maps will be able to provide more diversified associations for researchers. Another limitation is that this study only analyses simple associations between topics. The intensity of associations could be explored in further studies. Finally, conference information contains many useful knowledge resources for researchers. Existing research studies have suggested that conference information can reflect research trends and new issues (Palmer, 2013; Tomaszewski and MacDonald, 2009). In future studies, researchers may further explore research trends in their topics of interest using topic maps and by integrating conference information as their data resource.

References

Acedo, F.J., Barroso, C., Casanueva, C. and Galán, J.L. (2006), "Co-authorship in management and organizational studies: an empirical and network analysis", *Journal of Management Studies*, Vol. 43 No. 5, pp. 957-983.

Al-Rajebah, N.I. and Al-Khalifa, H.S. (2010), "Semantic relationship extraction and ontology building using wikipedia: a comprehensive survey", *International Journal of Computer Applications*, Vol. 12 No. 3, pp. 6-12.

Biswas, B., Nayak, V. and Shakya, H.K. (2014), "Comparison of algorithms for social networks using ontology", *International Journal of Computer Applications*, Vol. 85 No. 13, pp. 30-34.

Brank, J., Grobelnik, M. and Mladenic, D. (2005), "A survey of ontology evaluation techniques", *Proceedings of the Conference on Data Mining and Data Warehouses (SiKDD), Ljubljana, Slovenia*, pp. 166-170.

Chiu, D.Y. and Pan, Y.C. (2014), "Topic knowledge map and knowledge structure constructions with genetic algorithm, information retrieval, and multi-dimension scaling method", *Knowledge-Based Systems*, Vol. 67, pp. 412-428.

Cugmas, M., Ferligoj, A. and Kronegger, L. (2016), "The stability of co-authorship structures", *Scientometrics*, Vol. 106 No. 1, pp. 163-186.

Du, T.C., Li, F. and King, I. (2009), "Managing knowledge on the web: extracting ontology from HTML web", *Decision Support Systems*, Vol. 47 No. 4, pp. 319-331.

González-Albo, B. and Bordons, M. (2011), "Articles vs. proceedings papers: do they differ in research relevance and impact? A case study in the library and information science field", *Journal of Informetrics*, Vol. 5 No. 3, pp. 369-381.

Hamasaki, M., Matsuo, Y. and Takeda, H. (2007), "Ontology extraction using social network", paper presented at International Workshop on Semantic Web for Collaborative Knowledge Acquisition in Hyderabad.

Hu, C. and Racherla, P. (2008), "Visual representation of knowledge networks: a social network analysis of hospitality research domain", *International Journal of Hospitality Management*, Vol. 27 No. 2, pp. 302-312.

Hudson, J. (1996), "Trends in multi-authored papers in economics", *The Journal of Economic Perspectives*, Vol. 10 No. 3, pp. 153-158.

Hwang, S.Y., Wei, C.P. and Liao, Y.F. (2010), "Coauthorship networks and academic literature recommendation", *Electronic Commerce Research and Applications*, Vol. 9 No. 4, pp. 323-334.

Jiang, S.Q., Du, J., Huang, Q.M., Huang, T.J. and Gao, W. (2005), "Visual ontology construction for digitized art image retrieval", *Journal of Computer Science and Technology*, Vol. 20 No. 6, pp. 855-860.

Jiang, X. and Tan, A.H. (2009), "Learning and inferencing in user ontology for personalized semantic web search", *Information Sciences*, Vol. 179 No. 16, pp. 2794-2808.

Kim, J.M., Shin, H. and Kim, H.J. (2007), "Schema and constraints-based matching and merging of topic maps", *Information Processing & Management*, Vol. 43 No. 4, pp. 930-945.

Lee, C.S., Kao, Y.F., Kuo, Y.H. and Wang, M.H. (2007), "Automated ontology construction for unstructured text documents", *Data & Knowledge Engineering*, Vol. 60 No. 3, pp. 547-566.

Liu, X., Bollen, J., Nelson, M.L. and Van de Sompel, H. (2005), "Co-authorship networks in the digital library research community", *Information Processing & Management*, Vol. 41 No. 6, pp. 1462-1480.

Numprasertchai, S. and Igel, B. (2005), "Managing knowledge through collaboration: multiple case studies of managing research in university laboratories in Thailand", *Technovation*, Vol. 25 No. 10, pp. 1173-1182.

Oliva, J., Serrano, J.I., del Castillo, M.D. and Iglesias, Á. (2011), "SyMSS: a syntax-based measure for short-text semantic similarity", *Data & Knowledge Engineering*, Vol. 70 No. 4, pp. 390-405.

Palmer, C.L. (2013), *Work at the Boundaries of Science: Information and the Interdisciplinary Research Process*, Springer, New York, NY.

Pepper, S. (2010), "Topic maps", *Encyclopedia of Library and Information Sciences*, Taylor and Francis, Boca Raton.

Ponzetto, S.P. and Strube, M. (2011), "Taxonomy induction based on a collaboratively built knowledge repository", *Artificial Intelligence*, Vol. 175 Nos 9/10, pp. 1737-1756.

Sánchez, D. and Moreno, A. (2008), "Learning non-taxonomic relationships from web documents for domain ontology construction", *Data & Knowledge Engineering*, Vol. 64 No. 3, pp. 600-623.

Santoso, H.A., Haw, S.C. and Abdul-mehdi, Z.T. (2011), "Ontology extraction from relational database: Concept hierarchy as background knowledge", *Knowledge-Based Systems*, Vol. 24 No. 3, pp. 457-464.

Shah, P.K., Perez-Iratxeta, C., Bork, P. and Andrade, M.A. (2003), "Information extraction from full text scientific articles: where are the keywords?", *BCM Bioinformatics*, Vol. 4 No. 1, pp. 20-28.

Slawsky, D. (2007), "Building a keyword library for description of visual assets: thesaurus basics", *Journal of Digital Asset Management*, Vol. 3 No. 3, pp. 130-138.

Tomaszewski, R. and MacDonald, K.I. (2009), "Identifying subject-specific conferences as professional development opportunities for the academic librarian", *The Journal of Academic Librarianship*, Vol. 35 No. 6, pp. 583-590.

Venkatesh, V., Shaw, S., Dicks, D. and Lowerison, G. (2007), "Topic maps: adopting user-centred indexing technologies in course management systems", *Journal of Interactive Learning Research*, Vol. 18 No. 3, pp. 429-450.

Vigneshwari, S. and Aramudhan, M. (2015), "Social information retrieval based on semantic annotation and hashing upon the multiple ontologies", *Indian Journal of Science and Technology*, Vol. 8 No. 2, pp. 103-107.

Vrandečić, D. (2009), "Ontology evaluation", in Staab, S. and Studer, R. (Eds), *Handbook on Ontologies*, Springer, Berlin, pp. 293-313.

Wang, H.C., Huang, T.H., Guo, J.L. and Li, S.C. (2009), "Journal article topic detection based on semantic features"*International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems in Tainan, Taiwan, Springer, Berlin*, pp. 644-652.

Wu, Y., Lehman, A.R. and Dunaway, D.J. (2015), "Evaluations of a large topic map as a knowledge organization tool for supporting self-regulated learning", *Knowledge Organization*, Vol. 42 No. 6, pp. 386-398.

Xia, L., Wang, Z., Chen, C. and Zhai, S. (2016), "Research on feature-based opinion mining using topic maps", *The Electronic Library*, Vol. 34 No. 3, pp. 435-456.

Yao, Y., Lin, L., Wang, F. and Zhang, W. (2013), "Multi-perspective modeling: managing heterogeneous manufacturing knowledge based on ontologies and topic maps", *International Journal of Production Research*, Vol. 51 No. 11, pp. 3252-3269.

Yi, M. (2008), "Information organization and retrieval using a topic maps-based ontology: results of a task-based evaluation", *Journal of the American Society for Information Science and Technology*, Vol. 59 No. 12, pp. 1898-1911.

Zhai, C. and Massung, S. (2016), *Text Data Management and Analysis: A Practical Introduction to Information Retrieval and Text Mining*, Association for Computing Machinery, New York, NY.

About the authors

Hei Chia Wang is presently working as a Professor in the Institute of Information Management at National Cheng Kung University, Taiwan. His research focuses on knowledge discovery, text mining, e-learning and bioinformatics. Wang obtained both MSc in information system engineering and PhD in informatics from the University of Manchester (UMIST), UK. Hei Chia Wang is the corresponding author and can be contacted at: hcwang@mail.ncku.edu.tw

Yu Hung Chiang is a PhD Student in the Department of Industrial and Information Management Institute of Information Management at National Cheng Kung University, Taiwan. His advisor is Dr Hei-Chia Wang. Yu-Hung's main research focus on social networking service, e-learning and information retrieval.

Yen Tzu Huang has graduated from the Department of Industrial and Information Management Institute of Information Management at National Cheng Kung University, Taiwan. Her advisor is Dr Hei-Chia Wang.