Scientometrics (2015) 105:2035–2056
DOI 10.1007/s11192-015-1742-z

A multi-faceted method for science classification schemes (SCSs) mapping in networking scientific resources

Wei Du[1] · Raymond Yiu Keung Lau[1] · Jian Ma[1] · Wei Xu[2]

Received: 11 May 2015 / Published online: 16 September 2015

Abstract Science classification schemes (SCSs) are built to categorize scientific resources (e.g. research publications and research projects) into disciplines for effective research analytics and management. With the explosive growth of the number of scientific resources in distributed research institutions in recent years, effectively mapping different SCSs, especially heterogeneous SCSs that categorize different kinds of scientific resources, is becoming an increasingly challenging problem for facilitating information interoperability and networking scientific resources. To effectively realize the heterogeneous SCSs mapping, we design a novel multi-faceted method to measure the similarity between two classes based on three important facets, namely descriptors, individuals, and semantic neighborhood. Particularly, the proposed approach leverages a hybrid method combining statistical learning, semantic analysis and structure analysis for effective measurement with the exploitation of symmetric Tversky's index, WordNet dictionary and the Hungarian Algorithm. The method has been evaluated based on two main SCSs that need mapping for information management and policy-making in NSFC, and shown satisfying results. The interoperability among heterogeneous SCSs is resolved to enhance the access to heterogeneous scientific resources and the development of appropriate research analytics policies.

Keywords Science classification scheme (SCS) · Multi-faceted mapping · Semantic analysis · Research management

✉ Wei Du
weidu7-c@my.cityu.edu.hk

[1] Department of Information Systems, College of Business, City University of Hong Kong, Tat Chee Avenue, Kowloon, Hong Kong, Sar

[2] School of Information, Renmin University of China, No.59, Zhongguancun Street, Haidian District, Beijing, China

Introduction

Science classification schemes (SCSs), often represented as hierarchical structures of classes, have been widely used to categorize various scientific resources (e.g. research publications and projects) into disciplines (i.e. subjects or classes) and sub-disciplines based on their research areas. By representing the information resources with a neat structure, long-standing classification schemes have become a useful tool for information management such as information query, efficient information browsing and decision-making (Sokal 1974). First, the simple structure makes it amenable to computer-assisted retrieval (Chan 2000; Sokal 1974): class nodes provide a convenient access to detailed information thereby improving information retrieval efficiency. Second, the hierarchy structure makes it flexible in decision-making applications (Antelman et al. 2013; Tijssen 1992): attributes inheritance (i.e. subclasses and individuals inherit the attributes of their classes) enables the applications operating on different level of information granularity. Typical classification schemes for scientific resources such as web of science (WoS) subject categories categorizing academic journals and National Science Foundation discipline tree organizing research projects also play critically important role in research management (Fall et al. 2003; Larkey 1999; Rowley 1994). Scientometrics and Bibliometrics research utilize SCSs to detect disciplinary structure, identify emerging subfields, and analyze relations between different research fields (Boyack et al. 2005; Noyons 2001; Noyons et al. 1999; Szostak 2008; Vugteveen et al. 2014), while research institutions apply SCSs to facilitate research analytics and management (Rafols and Leydesdorff 2009). Besides, SCS with predetermined and uniform classes has advantages in research analytics over word or citation analysis (Tijssen 1992). For the purpose of performance evaluation of research entities, SCSs are often rebuilt or subdivided to reduce the field bias by classifying them into appropriate research fields (Glänzel and Schubert 2003; Glänzel et al. 1999; Noyons 2001; Noyons et al. 1999; Robinson-García and Calero-Medina 2013).

In reality, it is always difficult to build an ideal and universally accepted SCS. For information management not limited in scientific area, different institutions are accustomed to design and apply their own classification schemes for distinct purposes (Pfeffer 2014). Improving the interoperability and networking among the distributed and increasingly massive information has become a critical and challenging problem (Lei Zeng and Mai Chan 2004; Omelayenko 2002). Accordingly, in scientific area, different SCSs organized scientific resources tend to become isolated information islands, and cause great difficulties in information sharing and networking among different research institutions such as funding agencies, bibliographic databases and universities. Due to this, it's necessary to realize mapping between different SCSs by providing most associated class pairs.

This research takes the case of NSFC (National Science Foundation of China) that needs communication with other research institutions for research funding management. As one of the largest funding agencies in China, NSFC funds about 16, 000 general research projects with about RMB 10,000 million[1] every year as shown in Table 1 and organizes them according to NSFC discipline tree (i.e. SCS). Other scientific information related with research projects in NSFC are organized by different SCSs, for example, journal publications (about an annual output of 80,000 papers as shown in Table 1) are organized in bibliographic databases such as Thomson Reuters web of science (WoS) and China National Knowledge Infrastructure (CNKI) library according to their research fields.

[1] http://www.nsfc.gov.cn/nsfc/cen/xmtj/

Table 1 Scientific resources organized in NSFC

Year	The number of *approved general research projects* (total budget)	The number of *finished general research projects*	Output of *finished general research projects*	
			International journal publications	Local journal publications
2011	15,329 (RMB 8989.41 million)	7666	16,908	58,303
2012	16,891 (RMB 12, 480 million)	9031	39,771	43,160
2013	16,194 (RMB 12, 000 million)	9984	47,264	43,962

Considering the huge amount of distributed information, it's extremely useful for NSFC managers to know which journal subject categories are most related with a given NSFC discipline when retrieving relevant journals, or finding potential reviewers who own the most relevant publications. Semantic relations between classes of research publications and NSFC research projects are also expected to efficiently facilitate the process of research management applications such as recommendation systems (Xu et al. 2012) and reviewer assignment (Silva et al. 2013) for NSFC. Last, mapping between classes provide convenience for bilingual scientific information processing with the easier translation of class descriptors in comparison to translation of detail information such as full-text research publications and research proposals.

Existing research about classification scheme mapping is mainly conducted under homogeneous environment that deals with the same kind of entities, for example, mapping between library classification schemes has been extensively studied for networking library information (Chan 2000; Chaplan 1995; McCulloch et al. 2005; Pfeffer 2014; Zhang et al. 2011). To the best of our knowledge, there are very few studies examining the mapping problem of heterogeneous SCSs that categorize different kinds of scientific resources to link distribute scientific information. In addition, conventional mapping methods have difficulty in capturing semantic relations between two different scientific classes and leveraging hierarchical relations within scheme structure when applied directly in heterogeneous scientific environment.

To bridge the gap, we propose a multi-faceted mapping approach in this paper to address the interoperability problem of heterogeneous SCSs faced by various research institutions. In the proposed method, symmetric Tversky's index, WordNet similarity and Hungarian Algorithm are applied to profile the mapping degree between two classes in different SCSs from three facets: descriptors, individuals, and neighborhoods of two classes. The multi-faceted mapping method integrates both content and structure analysis to conduct semantic mapping between classes. Content analysis aims to profile the relations between two classes in terms of features such as their descriptors and individuals; structure analysis aims to profile the relations by leveraging the hierarchical relations among classes within SCS structure. In this research, the structure analysis compares not only the depths of two classes in their own SCSs but also their neighbors (e.g. super-classes, sub-classes). The proposed method can also be applied to perform dynamic temporal mappings of heterogeneous SCSs in different periods.

To evaluate the proposed multi-faceted mapping method, we selected two main SCSs that need mapping for research management in NSFC: the NSFC discipline tree for

research projects and the Thomson Reuters WoS subject categories for research publications. For NSFC, information mapping from researchers' research publications into appropriate NSFC discipline tree is necessary for management issues such as seeking for external reviewers, recommending research opportunities to potential researchers and determining journal list in each NSFC discipline for performance evaluation of research projects. This research provides the visualized mapping pattern between selected SCSs. The mapping results generated by the proposed method are then evaluated by domain experts and objectively assessed by precision and recall rate. Our evaluation based on NSFC data confirms that the proposed method performs well in comparing with other baseline mapping methods in mapping heterogeneous SCSs. Finally, we discuss the potential application of the proposed multi-faceted mapping method in accessing external scientific resources as well as research analytics and management at NSFC.

The rest of the paper is organized as follows. "Literature review" section summarizes previous research about classification scheme mapping methods as well as other mapping methods of knowledge structures not limited to scientific area. The proposed multi-faceted mapping method is illustrated in "A multi-faceted mapping method for heterogenous SCSs" section. "Evaluation and results" section provides the evaluation of the proposed method in NSFC. "Discussion" section discusses the potential application of the proposed mapping method in real-world research analytics and networking. The paper then concludes with a discussion of the directions for future research and improvements.

Literature review

This section reviews the mapping among classification schemes as well as mapping methods but not limited to scientific area, which lays the foundation of the proposed mapping method. Mapping among classification schemes is definitely not new. There have been various research studying mapping for different purposes. Generally, the mapping between two hierarchical structures aims to find corresponding or similar concepts through various similarity measures.

Two main objectives of classification scheme mappings are: (1) to realize interoperability across distributed information sources (Chan 2000; McCulloch et al. 2005), and (2) to improve information retrieval efficiency by integrating different classification schemes (Omelayenko 2002; Pfeffer 2014). Besides, mapping between different classification schemes is also useful to evaluate and formalize existing classification schemes by reference to relatively formal ones, for example, three main library classification schemes are evaluated by human knowledge in pillars to examine the efficiency of existing library classifications schemes in covering human knowledge (Zins and Santos 2011).

To our knowledge, previous research about classification scheme mappings mainly follows two streams: one stream follows the traditional manual mapping, while another adopts statistical analysis for automatic or semi-automatic mapping. Conventional manual mapping relies mainly on the ability and experience of experts. It can be adopted as a relatively reliable method especially when there is not enough information about individuals of classes. For example, Laborline Thesaurus terms are manually mapped to LCSH (Library of Congress Subject Headings) (Chaplan 1995) with matching degree from 1 to 19. Manual mapping is feasible though labor-intensive. Compared to manual mapping, statistical analysis based automatic or semi-automatic mapping is more efficient. However, it's difficult for most methods to be applied automatically without human involvement.

Due to this, semi-automated mapping with human experts who manually set parameters and analyze internal structures of classification schemes is more reliable and controllable (Lei Zeng and Mai Chan 2004). To reduce the heterogeneity and improve sharing among segmented German library classifications, Pfeffer (2014) proposes a simple method to cluster entries from several library classification schemes by comparing generated keys of entries. With the resulting clusters, large numbers of previously not indexed entries can be indexed by sharing indexing and classification information, and the information retrieval efficiency is largely improved. To facilitate the process of business document transformation and document integration, Class2Class bridges are built based on Bayesian approach and co-occurrence of terms to map different product classification schemes in two databases (Omelayenko 2002). An automatic mapping system from DDC (dewey decimal classification) to CLC (Chinese library classification) is built for multi-lingual information retrieval, and the most relevant CLC number is selected for each DDC number according to frequency ranking of their overlapping documents (Zhang et al. 2011).

Mappings among other knowledge structures such as ontologies are also summarized in this paragraph as the foundation of our method. According to different information involved in computing similarity, there are mainly three kinds of mapping approaches (Kalfoglou and Schorlemmer 2003; Rahm and Bernstein 2001): content-based mapping, structure-based mapping and hybrid mapping. Content-based mapping approach defines the matching degree between two classes by comparing name-equality, overlapping instances or common knowledge domain through statistical analysis and semantic analysis (Breitman et al. 2008; Choi et al. 2006; Kalfoglou and Schorlemmer 2003; Marshall et al. 2006; Thor et al. 2007; Zins and Santos 2011). Most previous research about classification scheme mappings follows this approach. Structure-based mapping approach measures the matching between two classes by considering their neighbors' matching degree, or their distance (Avesani et al. 2005; Kalfoglou and Schorlemmer 2003). Both content-based and structure-based mapping approaches have defects. Due to this, by integrating the information in both content level and structure level, the hybrid mapping approach provides a systematic way to match entities from different hierarchical structures (Duong et al. 2009; Jiang and Conrath 1997; Rodríguez and Egenhofer 2003; Truong et al. 2013).

Most current mappings between classification schemes that organize same kind of entities (e.g. library documents, products and patents) and represent in similar structure are known as homogeneous classification scheme mapping. Mapping between heterogeneous classification schemes that organize different kinds of entities, especially in scientific area

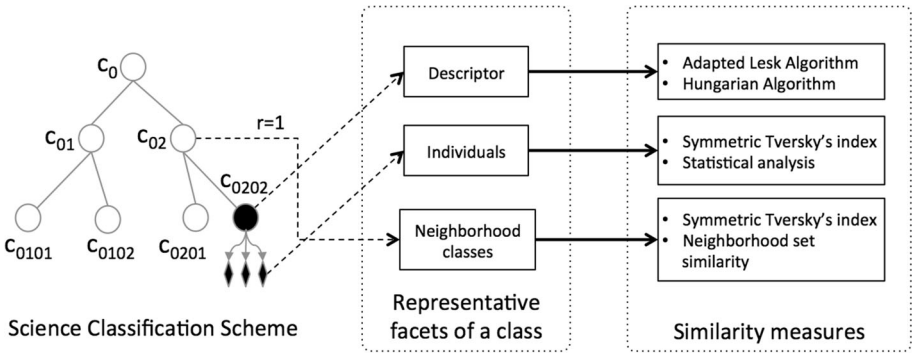


Fig. 1 Similarity measurement for heterogeneous SCSs

for networking scientific resources, are few. Mapping approaches summarized above can be used as the foundation of heterogeneous SCS mapping. Still, other problems such as semantic relations between scientific terms remain unsolved due to the lack of a standard taxonomy/ontology for the whole scientific area (Batet et al. 2011; Du et al. 2014; Gao and Xu 2013). Therefore, we combine WordNet dictionary and optimal matching technique in this research to measure semantic relations between scientific terms. Besides, features such as individuals and neighborhood classes are also considered in our method.

A multi-faceted mapping method for heterogenous SCSs

As shown in Fig. 1, three representative facets (i.e. features) of a class in a SCS are: its descriptor comprised of scientific terms or several sentences that describe the aim and scope of a class, its individuals that belong to the research field of a class in reality, and its neighborhood classes (e.g. sub-classes or super-classes) that have similar features. According to Tversky's claim (Hossein Zadeh and Reformat 2013; Tversky 1977), similarity between two objects can be profiled by the similarity between their features. Therefore, we propose a multi-faceted method capturing the semantic relations from descriptors, individuals, and the neighborhood of two classes. Semantic descriptor similarity measure integrates the adapted Lesk Algorithm (Banerjee and Pedersen 2003) and Hungarian Algorithm (Kuhn 1955; Melnik et al. 2002) to measure the relatedness between descriptors of two scientific classes based on WordNet dictionary. Besides, we use symmetric Tversky's index to balance content and structure information to measure the individual similarity and semantic neighborhood similarity. Time is also considered in the proposed method. Let $SCS_1$and SCS_2 represent two sets of classes from different SCSs. $c_1 \in SCS_1$ and $c_2 \in SCS_2$ denote two classes in the respective SCSs. The mapping relation between $SCS_1$and SCS_2 in a given time period T can be denoted as R^T (SCS_1, SCS_2).

Symmetric Tversky's index

In hierarchical schemas, features of a sub-class and its super-class always have different contributions when measuring their similarity (Tversky 1977). For example, according to Tversky's claim, the similarity of a sub-class c_{0202} to its super-class c_{02} is always smaller than the similarity of c_{02} to c_{0202} because $|I^{c_{02}}| > |I^{c_{0202}}|$ where $|I^{c_{02}}|$ denotes the number of individuals of c_{02}. Tversky's index has been widely used in measuring asymmetric similarity between two hierarchical schemas for its advantage in profiling the different contribution (Hossein Zadeh and Reformat 2013; Jimenez et al. 2013; Rodríguez and Egenhofer 2003). Tversky's index (Tversky 1977) is represented by:

$$S(A,B) = \frac{|A \cap B|}{|A \cap B| + \alpha|A - B| + \beta|B - A|} \quad (1)$$

where $|A \cap B|$ denotes the cardinality of common features between A and B, $A - B$ denotes the relative complement of B in A, and $\alpha, \beta \geq 0\,(\alpha + \beta = 1)$ are parameters of the Tversky's index. Tversky's index not only considers the common features, but depends also on features that are unique to each object. By using different parameters α and β to denote the relative importance of $|A - B|$ and $|B - A|$, the asymmetric similarity can be established. The different selection of A or B as referent outputs different similarities: $S(A, B) \neq S(B, A)$.

However, there are two main drawbacks of Tversky's index when it is applied in measuring similarity between facets of two scientific classes. On the one hand, descriptor-matching similarity based on the Tversky's index is difficult to capture the semantic relations of scientific classes from two SCSs. The comparison of two bags of words or its synonyms is not enough to measure the semantic similarity between two scientific terms. On the other hand, Tversky's index is not symmetric and the parameters α and β have the dual interpretation of modeling the asymmetry in the referent selection, while controlling the balance between common features $|A \cap B|$ and non-common features $|A - B| + |B - A|$ (Jimenez et al. 2013). Due to this, symmetric Tversky's index (Jimenez et al. 2013) is introduced as the foundation of similarity measures in our approach:

$$\begin{aligned} \text{Symmetric Tversky's index}: S_{sti}(A,B) &= \frac{|A \cap B| + bias}{|A \cap B| + bias + \beta\{\alpha a + (1-\alpha)b\}} \\ a &= \min(|A - B|, |B - A|) \\ b &= \max(|A - B|, |B - A|) \end{aligned} \tag{2}$$

where a denotes the minimum difference between cardinalities of A and B and b denotes the maximum difference. Parameter α controls the relative weight between a and b, while β is responsible for balancing common features $|A \cap B|$ and non-common features $|A - B| + |B - A|$. $\alpha < 0.5$ means that the minimum difference a obtains less weight in comparison to the maximum difference b. $\beta < 1$ means that common features are viewed more important than non-common features in computing similarity. Minimum difference a and maximum difference b as shown in Eq. (2) enable the similarity symmetric. Parameter *bias* denotes the common features that are frequent but less informative, e.g. stop words.

Classes in high level are more general and own higher data capacity than their sub-ordinate classes in SCS structure. By using *Science* as the common root of the selected SCSs, the depths of classes in each classification scheme can be obtained. For example, the depth of *A01-Mathematics* under root node *Science* is 1, while the depth of its subclass *A0102-Algebra* is 2. Classes in the higher level of the hierarchical structure are more generic and contain more information, while classes in the lower level of the hierarchical structure are more specific and contain less information. Difference between the depths of two classes can be used to estimate their difference. Therefore, parameters α and β can be defined as follows according to the interpretation in symmetric Tversky's index:

$$\begin{aligned} \alpha &= \min\left(\frac{dep_A}{dep_A + dep_B}, \frac{dep_B}{dep_A + dep_B}\right) \\ \beta &= 1 - \frac{|A \cap B|}{|A \cup B|} \end{aligned} \tag{3}$$

where parameter $0 < \alpha \leq 0.5$ denotes the relative difference between $|A - B|$ and $|B - A|$, and can be represented by the difference between their depths. According to Eq. (2), $0 \leq \beta \leq 1$ is used to balance the importance of non-common features and common features.

Semantic descriptor similarity

We view descriptor of a class as a sequence of scientific words, and semantic descriptor similarity is transformed to comparing two sequences of scientific words. In this section, we adopt adapted Lesk Algorithm to measure word-pair similarity by using WordNet dictionary first, and then introduce Hungarian Algorithm to measure descriptor-pair similarity through optimal matching. We chose the adapted Lesk algorithm measuring the word-pair similarity for two reasons: (1) it succeeds in obtaining high accuracy in word sense disambiguation (WSD) in comparison with other measures (Patwardhan et al. 2003), which means it can assign more accurate meaning to a word in a given context; (2) it can measure relatedness of two concepts (or senses) across part of speech (POS) boundaries and exceed the limit of *is-a* relation (Pedersen et al. 2004). POS tagging is the process of identifying words as nouns, verbs, adjectives, adverbs, etc. WordNet is a large lexical database of English including nouns, verbs, adjectives and adverbs as well as their glosses (i.e. dictionary definitions) (Miller 1995; Pedersen et al. 2004). It organizes related concepts into synonym sets, and further connects them via a variety of lexical relations including *synonym, antonymy, is-a* and *part-of*. Majority of WordNet relations connect words within the same POS tags. The adapted Lesk algorithm measures the similarity between two concepts by comparing their glosses as well as glosses of their neighbors in WordNet, which enables the similarity measurement across POS boundaries possible.

For the descriptor of each class, stop words (e.g. *is, of, a, an, and,* etc.) are filtered out first. POS tags of remained words are obtained beforehand, and most of them are nouns and adjectives. We summarize the steps of measuring semantic descriptor similarity between two classes as follows.

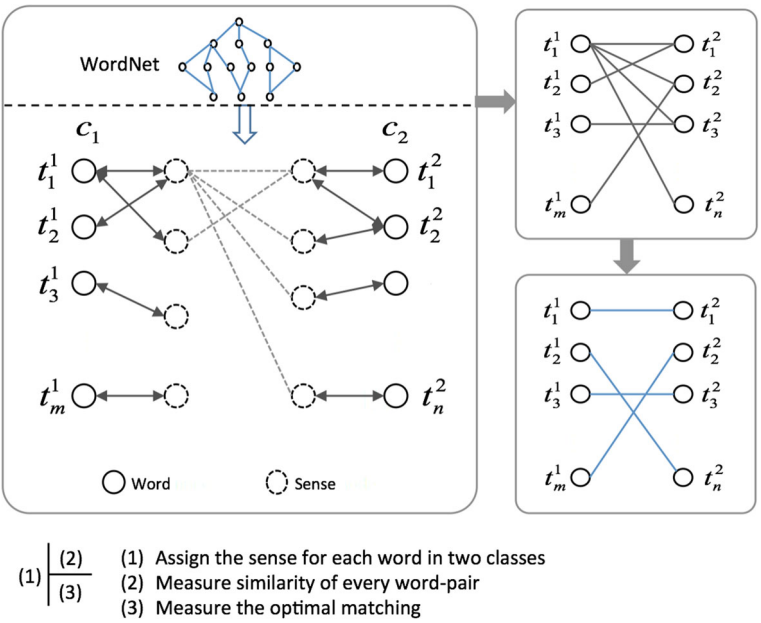


Fig. 2 WordNet based semantic descriptor similarity

Step 1 assign the sense for each word in two classes

One word may have several senses (i.e. meanings) in WordNet dictionary. For each word, the most appropriate sense in WordNet is assigned by using the adapted Lesk based Word Sense Disambiguation technique. The target word is assigned with the sense whose gloss has the most overlapping words with the glosses of its surrounding words. We name the number of distinct scientific words in the descriptor of a class after removing stop words as its descriptor-length. Suppose that the descriptor of class c_1 comprises a set of m distinct words: $\{t_1^1, t_2^1, \ldots, t_m^1\}$, then the descriptor-length of class c_1 is m. For each word t_i^1, the rest of words in the set are its surrounding words in the context of the scientific descriptor. Classc_2 comprises a set of n distinct words: $\{t_1^2, t_2^2, \ldots, t_n^2\}$. The obtained senses of two classes can be represented as: $c_1 = \{s_1^1, s_2^1, \ldots, s_m^1\}$ and $c_2 = \{s_1^2, s_2^2, \ldots, s_n^2\}$. This step can be described in Fig. 2-(1).

Step 2 measure the word-pair similarity

Word-pair similarity is translated into sense-pair similarity after Step 1. The adapted Lesk algorithm defines the relatedness rel_{ij} between sense s_i^1 and s_j^2 in WordNet by measuring the overlaps of their extended glosses (i.e. dictionary definitions in WordNet). Extended glosses include not only the gloss of the target sense but also those of the words that are related with the target sense through relations in WordNet. For example, the adapted Lesk relatedness between {*Optimization Theory*} and {*Operations Research, Management Science*} can be denoted by the matrix $\begin{pmatrix} 2 & 0 & 4 & 0 \\ 4 & 11 & 2 & 27 \end{pmatrix}$ where the first row denotes the WordNet-based relatedness between *Optimization* and each word in *Operations Research Management Science*. Relatedness values can be larger than 1. Therefore, we need to normalize the relatedness values into range of [0, 1] for similarity measurement. Similarity between two words is defined by normalizing the relatedness: if $rel_{ij} = 0$, $sim_{ij} = 0$; if $s_i^1 = s_j^2$, $sim_{ij} = 1$; otherwise, we use the log transformation method as one of the well-known normalization procedures (Attig and Perner 2011; Perner and Zotenko 2011), we define $sim_{ij} = \log_{\gamma \times rel_{\max}} rel_{ij}\ (\gamma > 1)$, where $rel_{\max}$is the maximum relatedness value of all word-pairs involved and γ parameter is introduced to normalize the relatedness values into range of (0, 1).

Step 3 measure the semantic descriptor similarity

Semantic descriptor similarity measurement can be transformed into weighted bipartite graph matching problem by viewing words in the descriptor of one class as vertices and the similarity between these words as the edges. Maximum weighted matching of the bipartite graph can be obtained through Hungarian Algorithm (Melnik et al. 2002). As shown in Fig. 2-(2), vertices denoting words are listed at both sides, and the length of each edge e_{ij} is defined as the similarity sim_{ij}. As shown in Fig. 2-(3), the optimal matching $E' \subseteq E$ from original matching $E$is to get the maximum value $\sum_{\forall e_{ij} \in E'} sim_{ij}$. For example, the resulted matched pairs of optimal bipartite matching between *Optimization Theory* and *Operations Research Management Science* is obtained as {*optimization-management*, *theory-science*}. The semantic descriptor similarity between two classes is measured by dividing the similarity summation of optimal matching by the maximum descriptor-length of two classes for normalization:

$$R(c_1, c_2, T)_{des} = \frac{\sum_{\forall e_{ij} \in E'} sim_{ij}}{\max(m, n)} \tag{4}$$

where m and n are descriptor-length values of two classes c_1 andc_2, and $T$is the selected time period for mapping SCSs.

Individual similarity

Different SCSs organize different research entities as individuals for each class. For two classes from different SCSs, if predefined relations exist between two of their individuals, one matched individual pair is identified. Overlapping individuals of two classes from two different SCSs can represent their common part in similarity measurement. Overlapping individuals vary with time: new individuals emerge and outdated individuals are excluded. Parameters α and β in symmetric Tversky's index are defined in Eq. (3). Thus, we define temporal individual similarity as:

$$\begin{aligned} R(c_1, c_2, T)_{ind} &= \frac{|S^T(I^{c_1}) \cap S^T(I^{c_2})|}{|S^T(I^{c_1}) \cap S^T(I^{c_2})| + \beta\{\alpha a + (1-\alpha)b\}} \\ a &= \min\left(\left|S^T(I^{c_1}) - S^T(I^{c_2})\right|, \left|S^T(I^{c_2}) - S^T(I^{c_1})\right|\right) \\ b &= \max\left(\left|S^T(I^{c_1}) - S^T(I^{c_2})\right|, \left|S^T(I^{c_2}) - S^T(I^{c_1})\right|\right) \end{aligned} \tag{5}$$

where $S^T(I^{c_1})$ denotes the set of individuals of the class $c_1 \in SCS_1$ during time period T, $S^T(I^{c_1}) \cap S^T(I^{c_2})$ is the set of all matched individual pairs between two classes$c_1 \in SCS_1$ and $c_2 \in SCS_2$, and $S^T(I^{c_1}) - S^T(I^{c_2})$ is the relative complement of c_2 in c_1 in terms of individuals.

Semantic neighborhood similarity

Semantic neighborhood similarity estimates the similarity between two classes by using the similarity between two sets of their neighbors. Classes are regarded as the neighbors of c_1 within $r$if the shortest distance between them in the hierarchical classification scheme is less than r. $r = 1$ is set to measure the shallow equality (Zdonik and Maier 1990), which means that only superclasses or subclasses of classes are collected to measure the semantic neighborhood similarity. According to Rodríguez and Egenhofer's work (Rodríguez and Egenhofer 2003), the semantic neighborhood similarity in time period T with distance r is defined based on symmetric Tversky's index:

$$\begin{aligned} &R(c_1, c_2, T, r)_{\text{neighbor}} = \frac{\left\|c_1 \cap_n^T c_2\right\|}{\left\|c_1 \cap_n^T c_2\right\| + \beta\{\alpha a + (1-\alpha)b\}} \\ &\text{where} \\ &a = \min\left(f\left(|c_1 - c_2|_n\right), f\left(|c_2 - c_1|_n\right)\right) \\ &b = \max\left(f\left(|c_1 - c_2|_n\right), f\left(|c_2 - c_1|_n\right)\right) \\ &f\left(|c_1 - c_2|_n\right) = |N(c_1, r)| - \left\|c_1 \cap_n^T c_2\right\| \end{aligned} \tag{6}$$

where $\|c_1 \cap {}_n^T c_2\|$ represents the approximate cardinality of the intersection set between two neighborhood sets, $N(c_1, r) = \{c_1^1, c_1^2, \ldots, c_1^i \ldots\}$denotes the cardinality of the

neighborhood set of class c_1 within distance r, $f(|c_1 - c_2|_n) \geq 0$ denotes the relative complement of c_2 in c_1 in terms of neighborhood set.

The approximate cardinality is defined by the relatedness between two neighborhood sets:

$$\left\|c_1 \cap_n^T c_2\right\| = \sum_{i \leq n} \max_{j \leq m} R_0^T(c_1^i, c_2^j) \tag{7}$$

where initial relatedness $R_0^T(c_1^i, c_2^j)$ can be estimated by weighted summation of descriptor similarity and individual similarity: $R_0^T(c_1^i, c_2^j) = w_d' R_{des}^T(c_1^i, c_2^j) + w_i' R_{ind}{}^T(c_1^i, c_2^j)$. The summation of weight values is equal to 1: $w_d' + w_i' = 1$. If relatedness values between the two neighborhood sets are all zero, we have $\|c_1 \cap {}_n^T c_2\| = 0$ and $R(c_1, c_2, T, r)_{neighbor} = 0$.

Similarity aggregation

The temporal similarity during time period T can be represented as the weighted aggregation of the three facets above, which is defined as follows:

$$R^T(c_1, c_2) = w_d R_{des}(c_1, c_2, T) + w_i R_{ind}(c_1, c_2, T) + w_n R_{neighbor}(c_1, c_2, T, r) \tag{8}$$

where similarity values of three facets are respectively normalized into [0, 1] for commensurability before the aggregation. w_d, w_i and w_n are the three weight values for similarity on semantic descriptor level, individual level and semantic neighborhood level respectively with the constraint of $w_d + w_i + w_n = 1$. Generally, weight values can be manually given by managers who are experienced in research management, which requires good domain knowledge of managers. An alternative method is to learn relatively optimal weight values through machine learning methods.

Evaluation and results

To evaluate the proposed multi-faceted mapping method, we realize the mapping between NSFC discipline tree for research projects and Thomson Reuters WoS subject categories for journal publications. We first give a brief introduction of selected SCSs and their relationships. Afterwards, weights of three facets are estimated through genetic algorithm to further obtain mapping results. We invite experienced managers in NSFC to manually determine the true mapping results. The performance of the proposed mapping method is examined thoroughly with any combination of facets in terms of precision and recall rate. Results show that the combination of three facets obtains best performance and the proposed method performs well in comparison with other similarity based mapping methods.

Data sets

In NSFC, NSFC discipline tree organizes the funded research projects while Thomson Reuters WoS subject categories and other SCSs organize the research output such as research publications and patents. NSFC builds eight research departments from *Mathematical and Physical Science Department* (*A*) *to Medical Science* (*H*) to organize research projects from various fields. A three-tier discipline tree under each department is built to further determine the specific academic area of research projects. Researchers can submit

proposals to discipline classes at various levels. Thomson Reuters WoS subject categories provides bibliographic and citation information of research publications. We select a total of 232 WoS subject categories as the combination of SCIE (Science Citation Index Expanded) and the SSCI (Social Science Citation Index) that cover journal publications in both sciences and social sciences field. By setting *Science* as the root node with depth = 0, the depths of classes in two selected SCSs can be determined according to "Symmetric Tversky's index" section.

The selected dataset includes research projects as well as attached information from seven selected departments (Medical Science excluded) in five years (i.e. from beginning of year 2009 to the end of year 2013) extracted from ISIS (Internet-based Science Information Systems)[2] for NSFC. The selected 79,106 general research projects are all finished in the period 2009–2013 and tagged with NSFC classes (1903 classes in total). Corresponding 1127,847 overall output research publications (where 443,528 of them are in WoS database) are collected. Data denoting the communications and relations between NSFC discipline tree and WoS subject categories are collected from the NSFC local database.

Mapping results

To measure the descriptor similarity, the parameter is set as $\gamma = 1.5$ in normalizing the word-pair similarity. For NSFC-WoS mapping, a matched individual-pair means a journal publication under a WoS category shares at least a researcher with a research project under a NSFC discipline class. Distance r in computing neighborhood similarity is set as 1, which means that only super-class and sub-classes are taken into account when computing the neighborhood similarity.

Weight values of the three facets are estimated through predefined true mappings and machine learning techniques. To obtain the estimation, we invited three experienced managers in NSFC to provide the most relevant NSFC-WoS pairs as true mappings since they are more familiar with the output distribution of research projects. During the process, experts found the assignment of relevant WoS classes to third level G-discipline codes are more difficult in comparison to high-level discipline codes. Therefore, to reduce their work and ensure the efficiency, they only needed to assign the most relevant WoS classes to the 42 selected discipline codes (first- and second-level discipline codes with finished research projects >1) from *G-Management Science* department in NSFC. We collect the manual mapping as true results for training weight values. Fleiss' Kappa (Fleiss 1971) was introduced to measure inter-rater agreement of the three experts. By using online software (Geertzen 2012), we compute Fleiss' Kappa = 0.718. According to Landis and Koch's (1977) study, the three experts have at least *substantial agreement* ($0.61 \leq \text{Kappa} \leq 0.80$) during the determination at the first and second level mapping. Disagreements are resolved through discussion.

We introduce genetic algorithm (Davis 1991) to estimate the weight values of three facets to achieve near optimal performance. F-measure (Eq. 11), as a harmonic mean of precision and recall by combining the two evaluation metrics evenly (Banerjee and Pedersen 2003; Makhoul et al. 1999), is selected to measure the performance of weight values and profile the fitness function. Here the relevant items are the manually given WoS subject categories for NSFC discipline classes. A population comprises a set of items, and each item is a vector with three weight values. With a selected item, the aggregated

[2] http://isisn.nsfc.gov.cn/egrantweb/

similarity between any two classes from NSFC discipline tree and WoS subject categories can be computed. For each NSFC class, the most relevant WoS class can be obtained by ranking the similarity from high to low, and the most relevant WoS class is viewed as the retrieved mapping items if it gets similarity larger than the given threshold minimum value $\lambda_{\min}$. Thus, for each population, the set of F-measure values can be computed. The population is evolved to obtain higher F-measure values in each iteration through selection, crossover and mutation. Genetic algorithm terminates when: reaching fixed number of generation, or identifying the solution with satisfying performance. The result population is stored after the termination.

$$\text{precision} = \frac{|\{\text{relevent items}\} \cap \{\text{retrieved items}\}|}{|\{\text{retrieved items}\}|} \tag{9}$$

$$\text{recall} = \frac{|\{\text{relevent items}\} \cap \{\text{retrieved items}\}|}{|\{\text{retrieved items}\}|} \tag{10}$$

$$F = \frac{2 \cdot \text{ precision} \cdot \text{ recall}}{\text{precision} + \text{recall}} \tag{11}$$

For the NSFC-WoS mapping, we select an item from the result population generated by genetic algorithm with threshold minimum value as $\lambda_{\min} = 0.01$. Larger threshold minimum value may lead to smaller precision or smaller recall because fewer real relevant classes are retrieved. The selected item has the largest F value: F@1 = 0.7619, Precision@1 = 0.7619 and Recall@1 = 0.7619. Here, Precision@n denotes the precision rate when top n potential mapping items are retrieved for the given class. Recall rate will monotonically increase if n increase. Recall rate converges at top 3: Recall@3 = 0.8810. The corresponding weight values for the three similarity components are $w_d^1 = 0.5885$, $w_f^1 = 0.1769$ and $w_n^1 = 0.2346$. The rest of the NSFC-WoS mapping can be computed through the proposed mapping method.

Obtained relations between NSFC classes and WoS subject categories not only establish interoperability among distributed information organized by different SCSs, but also provide implications for policy-making in the NSFC. To better visualize the mapping pattern between NSFC and WoS, we group NSFC by departments to detect the distribution of WoS subject categories over the seven NSFC departments. For NSFC classes, only NSFC-WoS pairs with mapping degree: (1) *greater than or equal to* 0.1 and (2) *located at the top three* are included in the group as representative mapping pairs. Mapping pattern between NSFC department groups and WoS subject categories are shown in Fig. 3. Detail annotation is listed below Fig. 3. For NSFC managers, the most relevant department-WoS mapping pairs provide the main subject categories of research output of research projects in each department. We can easily obtain two conclusions from Fig. 3. First, it's not surprising that NSFC departments categorize WoS subject categories well, which means research projects belonging to the same NSFC department always output research publications in certain WoS subject categories. For *G-Management Science* department as a group, the most relevant WoS subject categories are *Operations Research and Management Science* (*with aggregated mapping degree 4.2*), *Management* (*2.4*), *Business Finance* (*2.2*), *Automatic Control Systems* (*1.6*), *Computer Science Information Systems* (*1.4*), etc. *WoS-Operations Research and Management Science* is also relevant with *F-Information Science department* with aggregated mapping degree = 3.8. The main corresponding WoS subject categories of a NSFC department provide a standard for performance evaluation of NSFC research projects because research publications are one of the most important

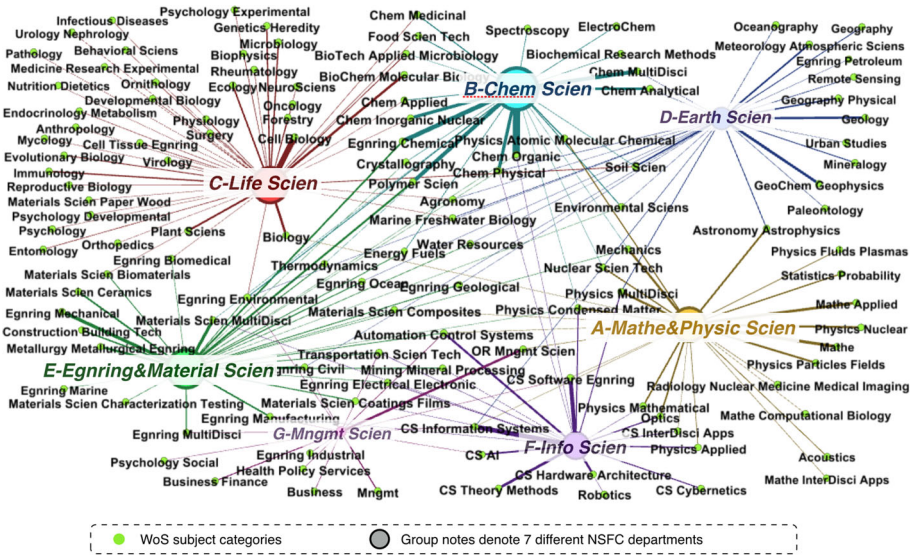


Fig. 3 Mapping between NSFC departments and WoS subject categories

features for research projects evaluation. Second, WoS subject categories link different NSFC departments, which reveals the interdisciplinarity across departments. For example, *G-Management Science* department relates with other departments (e.g. *F-Information Science*, *E-Engineering Material Science* and *A-Mathematics and Physics Science*) since they output research publications in common WoS subject categories (e.g. *WoS-Operations Research and Management Science, WoS-Engineering Multidisciplinary*).

Annotation: The seven bigger nodes with different colors in Fig. 3 denote the grouped discipline classes from the same NSFC department. We use the grouped discipline classes from the same department to denote the NSFC department to better reveal the relations between the NSFC departments and WoS subject categories. The size of each grouped node indicates the number of included subordinate representative classes in corresponding NSFC department: A-Mathematical and Physical Science Department includes 101 discipline classes, B-Chemistry Science Department includes 125 discipline classes, C-Life Science Department includes 106 discipline classes, D-Earth Science Department includes 65 discipline classes, E-Engineering and Material Science Department includes 114 discipline classes, F-Information Science Department includes 79 discipline classes and G-Management Science Department includes 30 discipline classes in Fig. 3. Other small green nodes distributed over Fig. 3 denote WoS subject categories. The link between each NSFC department and WoS subject category actually denotes the integrated association between subordinate NSFC discipline classes and the WoS subject category. Color of each link is the same with the color of the NSFC department node that belongs to the link, and by doing this we can easily tell the distribution of WoS subject categories over the seven departments. The weight of the link indicates the aggregated mapping degree between NSFC department and WoS subject category. For example, the link between *G-Management Science* department and WoS-*Operations Research and Management Science* denotes their relevance with aggregated mapping degree = 4.2.

To better show the mapping between the two selected SCSs in detail, we extract part of the mapping between discipline classes from G department in NSFC and corresponding WoS subject categories as shown in Fig. 4. For NSFC classes, only NSFC-WoS pairs with mapping degree greater than or equal to 0.1 are presented in Fig. 4 for clarity. The size of each node indicates the number of individuals. Purple nodes denote discipline classes from G department, and green nodes denote the WoS subject categories. The weight of the link between two nodes indicates the computed mapping degree according to our method and has been labeled beside each link. We can easily observe the most relevant WoS subject categories for each discipline class. For example, for G01-Management Science and Engineering class, the most relevant WoS subject categories with mapping degree $\geq$0.1 are *Operation Research and Management Science* (with mapping degree 0.2750), *Engineering Manufacturing* (0.1932), and *Computer Science Software Engineering* (0.1899). In addition, we can identify that discipline classes under the same superordinate class tend to link to the set of several WoS subject categories. This is consistent with the expectation that research contents of projects under the same or sibling discipline are similar, and thereby the output publications are mainly categorized in several similar fields according to WoS SCS. For example, the main corresponding WoS subject categories for the set of discipline classes under G01 are *Operations Research and Management Science*, Automation Control Systems and so on, while the main WoS subject categories for discipline classes under G02-Business Administration and G03-Macro Management *and* Policy are *Operations Research and Management Science, Business*, *Business Finance* and *Management*. Another interesting and significant application of the mapping is to revise the current SCS by using the relatively mature SCS. As shown in Fig. 4, all discipline classes under *G01-*

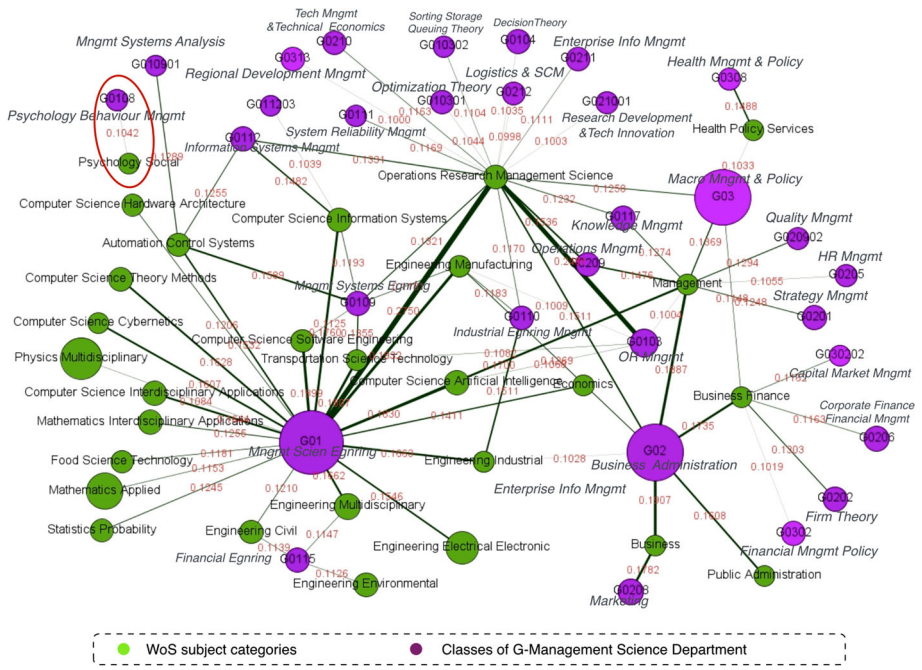


Fig. 4 Part of detail mapping between G-Management Science Department and WoS subject categories

Management Science and Engineering except *G0108-Management Psychology and Behavior* are connected indirectly via the common WoS subject category, which provides implication and evidence for NSFC managers to revise the scheme structure under G01. However, note that we only obtain the implications for improving SCSs by simple observation of Fig. 4 with cut-off value as 0.1. We still need to design a more delicate method for the potential application in revision and improvement of SCSs.

Comparative evaluation

In this section, we evaluate the proposed multi-faceted method by: (1) examining each facet and (2) comparing with other baseline methods for G-WoS mapping since we have the true mapping results from invited managers. Precision and recall are computed to measure the effectiveness of the proposed method and selected methods. As stated before, for the mapping between *G-Management Science* department and WoS subject categories (i.e. G-WoS), the optimal weight setting is ALL_optimal: (0.5885, 0.1769, 0.2346). To examine the importance of each facet, we conduct seven experiments by setting different weight values as shown in Table 2. The classic feature based similarity measures in Scientometrics research are selected as baseline methods: Jaccard Index (Boyack et al. 2005; Eck and Waltman 2009; Hamers et al. 1989), Cosine (Eck and Waltman 2009), Inclusion Index (Eck and Waltman 2009) and Association Strength (Boyack et al. 2005). The baseline methods are defined as:

$$\text{Jaccard Index}: R(c_1, c_2)_J = \frac{|S^T(I^{c_1}) \cap S^T(I^{c_2})|}{|S^T(I^{c_1}) \cup S^T(I^{c_2})|} \tag{12}$$

$$\text{Cos}ine: R(c_1, c_2)_C = \frac{|S^T(I^{c_1}) \cap S^T(I^{c_2})|}{\sqrt{|S^T(I^{c_1})| \times |S^T(I^{c_2})|}} \tag{13}$$

$$\text{Inclusion Index}: R(c_1, c_2)_I = \frac{|S^T(I^{c_1}) \cap S^T(I^{c_2})|}{\min(|S^T(I^{c_1})|, |S^T(I^{c_2})|)} \tag{14}$$

Table 2 Different weight values for examining the importance of each feature

Selected facets	Descriptor w_d	Individual w_i	Neighborhood w_n
1. Descriptor	1	0	0
2. Individual	0	1	0
3. Neighbor	0	0	1
4. Descriptor + individual	0.5	0.5	0
5. Descriptor + neighbor	0.5	0	0.5
6. Individual + neighbor	0	0.5	0.5
7. All	1/3	1/3	1/3

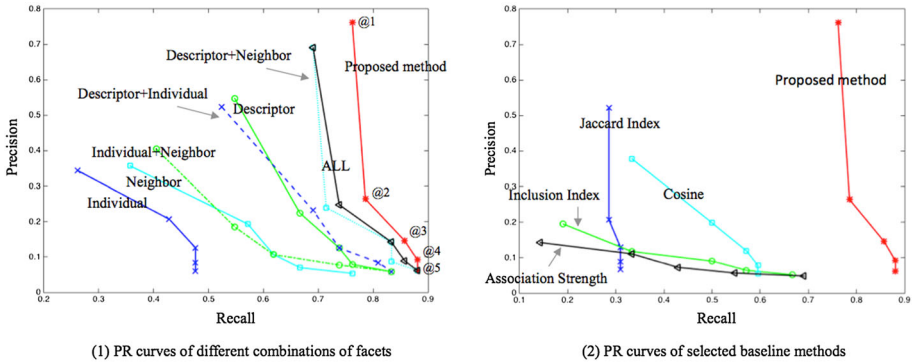
$$\text{Association Strength}: R(c_1, c_2)_A = \frac{|S^T(I^{c_1}) \cap S^T(I^{c_2})|}{|S^T(I^{c_1})| \times |S^T(I^{c_2})|} \quad (15)$$

To measure the mapping accuracy and compare the performance of the proposed method with other methods, we introduce precision and recall rate as defined in Eq. 9–11 that are widely used for both information retrieval and classification task (Ma et al. 2014; Yau et al. 2014). The performance of the proposed method and other methods are listed in Table 3 in terms of precision@1, recall@1 and F measure@1 with threshold minimum value $\lambda_{\min} = 0.01$. As we have mentioned in "Data sets" section, precision@n denotes the precision rate when top n potential mapping items are retrieved for the given class. Recall rate will monotonically increase if n increases because the intersection of the set of relevant items and the set of retrieved items expands. When measuring precision@1 and recall@1, the number of retrieved items (larger than threshold minimum value) can be very close to the number of relevant items. Therefore we got very close values of precision@1 and recall@1. Besides, to better show the performance of the proposed method, we also reveal values of precision and recall@1, 2, 3, 4 and 5 by plotting Precision-Recall (PR) curves (Boyd et al. 2013; Davis and Goadrich 2006) to visualize the effectiveness of each method. We compare the PR curve of the proposed method with the ones of different combinations of facets as shown in Fig. 5-(1), and with the ones of baseline methods as shown in Fig. 5-(2). Obviously, PR curve located at top right corner means the corresponding method performs better in mapping with relatively higher precision and recall values.

Results show that the proposed method performs well in comparison with other methods in terms of precision and recall. The reason why the precision and recall rate are not very high is because that there maybe more than one most relevant WoS subject categories for each NSFC. Results show that the proposed method with three involved facets performs better in mapping SCSs than the methods with one or two facets involved, while the method with optimal weight setting performs the best. Results also show the performance of each facet in mapping and the performance ranking is consistent with the importance in the optimal weight setting. For example, for the NSFC-WoS mapping, the weight ranking of the involved facets in decreasing order is Descriptor > Neighborhood > Individual in

Table 3 Precision, recall and F-measure@1 values for selected methods

Methods	Precision@1	Recall@1	F measure@1
Proposed method	0.7619	0.7619	0.7619
Descriptor	0.5476	0.5476	0.5476
Individual	0.3438	0.2619	0.2973
Neighbor	0.3571	0.3571	0.3571
Descriptor + individual	0.5238	0.5238	0.5238
Descriptor + neighbor	0.6905	0.6905	0.6905
Individual + neighbor	0.4048	0.4048	0.4048
ALL: descriptor + individual + neighbor (equal weights)	0.6905	0.6905	0.6905
Jaccard index	0.5217	0.2857	0.3692
Cosine	0.3784	0.3333	0.3544
Inclusion index	0.1951	0.1905	0.1928
Association strength	0.1429	0.1429	0.1429



(1) PR curves of different combinations of facets

(2) PR curves of selected baseline methods

Fig. 5 Precision-Recall (PR) curves of proposed method and selected methods

optimal weight setting, and the ranking of corresponding *F*-measure@1 of each single facet is the same. Figure 5-(1) also verifies the performance of each facet.

Discussion

In summary, the mapping results lead to several benefits for networking scientific resources therefore more accurate research analytics and research policy-making listed as follows.

1. Mapping facilitates information exchange among different research institutions. Based on the mapped class pairs produced by our method, accessing research entities across distributed information sources becomes possible. Thus, it improves information retrieval efficiency and facilitates information integration (Bergamaschi et al. 2001; Genesereth et al. 1997; Papakonstantinou et al. 1995).
2. The hidden relations among classes within the same SCS can also be inferred by an external SCS. For example, as shown in Fig. 4, *G0103-Opereations Research Management* and *G0117-Knowledge Management* are both subclasses of *G01*, which can be verified through their common relations with *WoS143-Operations Research Management Science*. Also, the *WoS143* class is relevant with other NSFC classes such as *A01-Mathematics* (*0.1607*) and *F0302- Systems Science and Systems Engineering* (*0.1366*). Their common relations with *WoS143* imply a possible similarity between these classes. Therefore, our mapping results enable the detection of relations/interdisciplinarity among classes within the same SCS in addition to the existing relations explicitly captured in SCS structure.
3. Mapping relations can be used to evaluate or revise existing SCSs. NSFC managers can apply WoS subject categories to evaluate the existing NSFC discipline tree by following some rules, for example, the NSFC classes close to each other tend to link common WoS subject categories more likely. Moreover, the interdisciplinarity of existing NSFC classes can be detected by referring to WoS subject categories.
4. The visualized relations among the backbones of distributed databases significantly enhance the networking of scientific resources therefore research management applications. In NSFC, one important criterion for evaluating research projects is research publications. For each NSFC class, the corresponding most relevant WoS subject categories provide good references of research field in comparing the research

publications, which is expected to generate a fairer evaluation of research projects (Glänzel and Schubert 2003). Besides, the corresponding most relevant WoS subject categories also provide a clue for finding potential external reviewers in a given NSFC discipline.

Conclusions and future work

SCSs, as the backbones of distributed scientific information sources, play a significant role in research management and have been applied to a variety of research analytics applications. However, isolated SCSs cause great difficulties in information communication and management across heterogeneous scientific resources. To bridge the gap between heterogeneous SCSs and achieve better interoperability and communications for networking information, we develop a novel multi-faceted mapping method to link the scientific classes captured in different SCSs. One novelty of the proposed method is that a multi-faceted mapping approach that leverages symmetric Tversky's index, WordNet based similarity, and the Hungarian Algorithm technique is applied to estimate the relations based on three main facets, namely descriptors, individuals, and neighborhood. Evaluation of the proposed method has been performed based on two SCSs adopted in NSFC by using a large number of scientific resources archived in NSFC. Our results confirm that the proposed method performs well when mapping a large number of classes between heterogeneous SCSs.

The practical implication of our research is that the mapping results, generated by the proposed mapping method, greatly facilitate the interoperability among heterogeneous SCSs and hence improve information retrieval as well as information integration across distributed scientific information sources. In addition, research institutions can leverage the visualized mapping patterns among various SCSs to support research analytics and decision-making.

Still, there are several limitations of our research presented in this paper. First, only the mapping results of a single time window are explored. Future work will incorporate the analysis of temporal mapping patterns by using multiple time windows. Second, we will perform a specific real-world research management application such as reviewer assignment system by using the networked SCSs to further test the proposed mapping method. Last but not least, the proposed mapping method will be incorporated into an operating research management system such as the online ScholarMate[3] platform for NSFC to obtain further insights about its benefits.

Acknowledgments The authors gratefully thank the Editor and all reviewers. The authors also acknowledge with gratitude the generous support of the University Grants Committee (UGC) of Hong Kong (CityU 148012), National Natural Science Foundation of China (71501057), and City University of Hong Kong.

References

Antelman, K., Lynema, E., & Pace, A. K. (2013). Toward a twenty-first century catalog. *Information Technology and Libraries, 25*(3), 128–139.

[3] http://www.scholarmate.com/

Attig, A., & Perner, P. (2011). The problem of normalization and a normalized similarity measure by online data. *Tran CBR, 4*(1), 3–17.

Avesani, P., Giunchiglia, F., & Yatskevich, M. (2005). A large scale taxonomy mapping evaluation. *Proceeding ISWC'05. Proceedings of the 4th international conference on The Semantic Web*, pp. 67–81.

Banerjee, S., & Pedersen, T. (2003). Extended gloss overlaps as a measure of semantic relatedness. *Ijcai, 3*, 805–810.

Batet, M., Sánchez, D., & Valls, A. (2011). An ontology-based measure to compute semantic similarity in biomedicine. *Journal of Biomedical Informatics, 44*(1), 118–125.

Bergamaschi, S., Castano, S., Vincini, M., & Beneventano, D. (2001). Semantic integration of heterogeneous information sources. *Data and Knowledge Engineering, 36*(3), 215–249.

Boyack, K. W., Klavans, R., & Börner, K. (2005). Mapping the backbone of science. *Scientometrics, 64*(3), 351–374.

Boyd, K., Eng, K. H., & Page, C. D. (2013). Area under the Precision-Recall Curve: Point estimates and confidence intervals. In *Machine learning and knowledge discovery in databases* (pp. 451–466). Springer.

Breitman, K. K., Brauner, D., Casanova, M. A., Milidiú, R., Gazola, A., & Perazolo, M. (2008). Instance-based ontology mapping. In *Engineering of Autonomic and Autonomous Systems, 2008. EASE 2008. Fifth IEEE Workshop on* IEEE, pp. 67–74.

Chan, L. M. (2000). Exploiting Lcsh, Lcc, and Ddc to retrieve networked resources: Issues and challenges.

Chaplan, M. A. (1995). Mapping laborline thesaurus terms to library of congress subject headings: Implications for vocabulary switching. *The Library Quarterly*, pp. 39–61.

Choi, N., Song, I.-Y., & Han, H. (2006). A survey on ontology mapping. *SIGMOD Record, 35*(3), 34–41.

Davis, L. (1991). *Handbook of genetic algorithms*. New York: Van Nostrand Reinhold.

Davis, J., & Goadrich, M. 2006. The Relationship between Precision-Recall and Roc Curves. *Proceedings of the 23rd international conference on Machine learning*: ACM, pp. 233–240.

Du, W., Xu, W., Jiang, H., & Ma, J. (2014). Fuzzy Classification Scheme Mapping for Decision Making. In *Thirty Fifth International Conference on Information Systems*. Auckland.

Duong, T. H., Nguyen, N. T., & Jo, G. S. (2009). A hybrid method for integrating multiple ontologies. *Cybernetics and Systems: An International Journal, 40*(2), 123–145.

Eck, N. J. V., & Waltman, L. (2009). How to normalize cooccurrence data? An analysis of some well-known similarity measures. *Journal of the American Society for Information Science and Technology, 60*(8), 1635–1651.

Fall, C. J., Törcsvári, A., Benzineb, K., & Karetka, G. (2003). *Automated categorization in the international patent classification* (pp. 10–25). ACM SIGIR Forum: ACM.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin, 76*(5), 378–382.

Gao, W., and Xu, T. (2013). Stability analysis of learning algorithms for ontology similarity computation. *Abstract and Applied Analysis, 2013*. doi:10.1155/2013/174802.

Geertzen, J. (2012). Inter-rater agreement with multiple raters and variables.

Genesereth, M. R., Keller, A. M., & Duschka, O. M. (1997). *Infomaster: An information integration system* (pp. 539–542). ACM SIGMOD Record: ACM.

Glänzel, W., & Schubert, A. (2003). A new classification scheme of science fields and subfields designed for scientometric evaluation purposes. *Scientometrics, 56*(3), 357–367.

Glänzel, W., Schubert, A., & Czerwon, H. J. (1999). An item-by-item subject classification of papers published in multidisciplinary and general journals using reference analysis. *Scientometrics, 44*(3), 427–439.

Hamers, L., Hemeryck, Y., Herweyers, G., Janssen, M., Keters, H., Rousseau, R., & Vanhoutte, A. (1989). Similarity measures in scientometric research: The Jaccard index Versus Salton's cosine formula. *Information Processing and Management, 25*(3), 315–318.

Hossein Zadeh, D., & Reformat, M. Z. (2013). Assessment of semantic similarity of concepts defined in ontology. *Information Sciences, 250*, 21–39.

Jiang, J. J., & Conrath, D. W. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. *arXiv preprint cmp-lg/9709008*).

Jimenez, S., Becerra, C., Gelbukh, A., Bátiz, A. J. D., & Mendizábal, A. (2013). *Softcardinality-core: Improving text overlap with distributional measures for semantic textual similarity*. p. 194. Atlanta, Georgia, USA.

Kalfoglou, Y., & Schorlemmer, M. (2003). Ontology mapping: The state of the art. *The Knowledge Engineering Review, 18*(01), 1–31.

Kuhn, H. W. (1955). The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly, 2*(1–2), 83–97.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159–174.
Larkey, L. S. (1999). A patent search and classification system. *Acm dl*, pp. 179–187.
Lei Zeng, M., & Mai Chan, L. (2004). Trends and issues in establishing interoperability among knowledge organization systems. *Journal of the American Society for Information Science and Technology, 55*(5), 377–395.
Ma, F.-C., Lyu, P.-H., Yao, Q., Yao, L., & Zhang, S.-J. (2014). Publication trends and knowledge maps of global translational medicine research. *Scientometrics, 98*(1), 221–246.
Makhoul, J., Kubala, F., Schwartz, R., & Weischedel, R. (1999). Performance measures for information extraction. *Proceedings of DARPA broadcast news workshop*, pp. 249–252.
Marshall, B., Chen, H., & Madhusudan, T. (2006). Matching knowledge elements in concept maps using a similarity flooding algorithm. *Decision Support Systems, 42*(3), 1290–1306.
McCulloch, E., Shiri, A., & Nicholson, D. (2005). Challenges and issues in terminology mapping: A digital library perspective. *The Electronic Library, 23*(6), 671–677.
Melnik, S., Garcia-Molina, H., & Rahm, E. (2002). Similarity flooding: A versatile graph matching algorithm and its application to schema matching. *Proceedings of the. 18th ICDE Conf.*(*Best Student Paper award*).
Miller, G. A. (1995). Wordnet: A lexical database for english. *Communications of the ACM, 38*(11), 39–41.
Noyons, E. (2001). Bibliometric mapping of science in a policy context. *Scientometrics, 50*(1), 83–98.
Noyons, E. C., Moed, H. F., & Van Raan, A. F. (1999). Integrating research performance analysis and science mapping. *Scientometrics, 46*(3), 591–604.
Omelayenko, B. (2002). Integrating vocabularies: Discovering and representing vocabulary maps. In *The Semantic Web—Iswc 2002*. , pp. 206–220. Springer.
Papakonstantinou, Y., Garcia-Molina, H., & Widom, J. (1995). Object exchange across heterogeneous information sources. *Data Engineering, 1995. Proceedings of the eleventh international conference on*: IEEE, pp. 251–260.
Patwardhan, S., Banerjee, S., & Pedersen, T. (2003). Using measures of semantic relatedness for word sense disambiguation. In *Computational linguistics and intelligent text processing*. Springer, pp. 241–257.
Pedersen, T., Patwardhan, S., & Michelizzi, J. (2004). Wordnet: Similarity: Measuring the relatedness of concepts. *Demonstration Papers at HLT-NAACL* 2004: Association for Computational Linguistics, pp. 38–41.
Perner, J., and Zotenko, E. (2011). *Characterizing Cell types through differentially expressed gene clusters using a model-based approach*. Springer.
Pfeffer, M. (2014). Using clustering across union catalogues to enrich entries with indexing information. In: *Data Analysis, Machine Learning and Knowledge Discovery*. pp. 437–445. Springer.
Rafols, I., & Leydesdorff, L. (2009). Content-based and algorithmic classifications of journals: Perspectives on the dynamics of scientific communication and indexer effects. *Journal of the American Society for Information Science and Technology, 60*(9), 1823–1835.
Rahm, E., & Bernstein, P. A. (2001). A survey of approaches to automatic schema matching. *The VLDB Journal, 10*(4), 334–350.
Robinson-García, N., & Calero-Medina, C. (2013). What do university rankings by fields rank? Exploring discrepancies between the organizational structure of universities and bibliometric classifications. *Scientometrics, 98*(3), 1955–1970.
Rodríguez, M. A., & Egenhofer, M. J. (2003). Determining semantic similarity among entity classes from different ontologies. *Knowledge and Data Engineering, IEEE Transactions on, 15*(2), 442–456.
Rowley, J. (1994). The controlled versus natural indexing languages debate revisited: A perspective on information retrieval practice and research. *Journal of information science, 20*(2), 108–118.
Silva, T., Guo, Z., Ma, J., Jiang, H., & Chen, H. (2013). A social network-empowered research analytics framework for project selection. *Decision Support Systems, 55*(4), 957–968.
Sokal, R. R. (1974). Classification: purposes, principles, progress, prospects. *Science, 185*(4157), 1115–1123.
Szostak, R. (2008). Classification, interdisciplinarity, and the study of science. *Journal of documentation, 64*(3), 319–332.
Thor, A., Kirsten, T., & Rahm, E. (2007). Instance-based matching of hierarchical ontologies. *BTW, 103*, 436–448.
Tijssen, R. J. W. (1992). A quantitative assessment of interdisciplinary structures in science and technology: Co-classification analysis of energy research. *Research Policy, 21*(1), 27–44.
Truong, H. B., Duong, T. H., & Nguyen, N. T. (2013). A hybrid method for fuzzy ontology integration. *Cybernetics and Systems, 44*(2–3), 133–154.
Tversky, A. (1977). Features of similarity. *Psychological Review, 84*(4), 327.

Vugteveen, P., Lenders, R., & Van den Besselaar, P. (2014). The dynamics of interdisciplinary research fields: The case of river research. *Scientometrics.*

Xu, Y., Guo, X., Hao, J., Ma, J., Lau, R. Y. K., & Xu, W. (2012). Combining social network and semantic concept analysis for personalized academic researcher recommendation. *Decision Support Systems, 54*(1), 564–573.

Yau, C.-K., Porter, A., Newman, N., & Suominen, A. (2014). Clustering scientific documents with topic modeling. *Scientometrics, 100*(3), 767–786.

Zdonik, S. B., & Maier, D. (1990). *Readings in object-oriented database systems*. Morgan Kaufmann.

Zhang, Y., Peng, J., Huang, D., & Li, F. (2011). Design of automatic mapping system between Ddc and Clc. In *Digital libraries: For cultural heritage, knowledge dissemination, and future creation* (pp. 357–366). Springer.

Zins, C., & Santos, P. L. (2011). Mapping the knowledge covered by library classification systems. *Journal of the American Society for Information Science and Technology, 62*(5), 877–901.