ATLANTIS PRESS International Journal of Computational Intelligence Systems, Vol. 10 (2017) 593–604

SAO Semantic Information Identification for Text Mining

Chao Yang, Donghua Zhu, Xuefeng Wang*

School of Management and Economics, Beijing Institute of Technology,
5 South Zhongguancun Street, Haidian District
Beijing 100081, China
E-mails: yc_2009@hotmail.com, zhudh111@bit.edu.cn, wxf5122@bit.edu.cn

Received 28 November 2016

Accepted 22 December 2016

Abstract

A Subject-Action-Object (SAO) is a triple structure which can be used to both describe topics in detail and explore the relationship between them. SAO analysis has become popular in bibliometrics, however there are two challenges in the identification of SAO structures: low relevance of SAOs to domain topics; and synonyms in SAO. These problems make the identification of SAO greatly dependent upon domain experts, limiting the further usage of SAO and influencing further the mining of SAO characteristics. This paper proposes a parse tree-based SAO identification method that includes (1) a model to identify the core components (candidate terms for subject & object) of SAO structures, where term clumping processes and co-word analysis are involved; (2) a parse tree-based hierarchical SAO extraction model to divide entire SAO structures into a collection of simpler sub-tasks for separate subject, action, and object identification; and (3) an SAO weighting model to rank SAO structures for result selection. The proposed method is applied to publications in the Journal of Scientometrics (SCIM), to identify and rank significant SAO structures. Our experiment results demonstrate the validity and feasibility of the proposed method.

Keywords: Semantic Analysis; Technology Intelligence; Computational Intelligence; Topic Model; Subject-Action-Object

1. Introduction

An SAO is a triple structure extracted from a text corpus. Subjects and objects are terms or phrases that are closely related to the topic. Actions are verbs that represent the operation by which, or the relationship between, those terms and phrases. The development of natural language processing techniques has allowed SAO structures to express rich semantic information and gained recognition as a powerful tool for identifying concepts in a corpus.[1-3] SAO structure has the ability to transform words into concepts,[4] express means-end relationships,[5] and also express evolutionary trends of topic.[6-8]

Compared to co-word analysis, SAOs provide an express way to quickly understand massive textual content. Co-word analysis, also named as co-occurrence analysis, is widely introduced to explore potential relationships between textual elements.[9-12] Undoubtedly, compared with simply counts of publications and citations, as described in the definition of modern bibliometrics,[13] co-words analysis provide a tool to quickly understand textual content,[14] and help indicate significant topics [15] and dynamic changes [16]. However, limitations still exist: (1) homonyms and synonyms of words and terms result in ambiguous interpretations, especially in multidisciplinary and interdisciplinary research fields [17, 18]; (2) High-frequency, or common, terms fail to discriminate between relevant and irrelevant topics and mislead to unexpected groups [15, 19]; and (3) co-occurrence only means two terms that appear at the same time, yet directly identifying the specific and deterministic relationships between them. Undoubtedly, compared to co-word analysis, SAOs

*Corresponding author.

Copyright © 2017, the Authors. Published by Atlantis Press.
This is an open access article under the CC BY-NC license (http://creativecommons.org/licenses/by-nc/4.0/).

provide an express way to quickly understand massive textual content, and help indicate significant topics. SAO is helpful for (1) solving the problem of ambiguous interpretations resulted by homonyms and synonyms of words [17, 18]; and (2) identifying the specific relationship between topic terms.[20]

SAO identification is the basis of SAO analysis. However, it is difficult to identify appropriate SAOs for bibliometric analysis. The problems of traditional SAO identification are: (1) it is difficult to directly extract the SAOs that have a close relationship with a topic of interest. Most of the SAOs identified with general Natural Language Processing are too common to express detailed meanings for "what to do" and "how to do it," which is the emphasis of "SAO structure" playing an important role in topic analysis. The reason is that there are millions of SAOs and most of them are common words, (e.g., "we look at an example," "paper consists of three parts."). These common words are irrelative to topics, stop us from getting truly valuable SAOs, and cannot be filtered out by post-cleaning and consolidation. (2) It is difficult to obtain the SAOs that have perfect quantitative properties and we usually face the problem of "synonyms in SAO". It is a serious problem for quantitative analysis, especially when we want to use some statistics-based methods (e.g., time series analyses,[21] co-occurrence and association analysis). The reason for "synonyms in SAO" is that the SAO structure is complex, which means that there will be literally many different SAO structures that have the same meaning, and it is difficult to combine them together.

Aiming to overcome the problems described above, which is the value of this manuscript, this paper proposes a SAO identification method. Compared with traditional SAO identification methods, the main contributions of the proposed method are: (1) introduce term clumping and design a co-word algorithm (considering the co-occurrence with keywords) to identify SAO core components, which is helpful for improving the relevance of SAOs to topic. (2) Based on syntax-tree, constructed a hierarchical SAO extraction model, and perform the SAO cleaning and consolidation function. It is helpful for improving the "synonyms in SAO". (3) Constructed an SAO weighting model using the idea of TFIDF (term frequency–inverse document frequency) to evaluate the importance of each SAO. We apply the proposed method to the publications in the *Journal of Scientometrics*. The results demonstrate the feasibility of our method and hold interest for related bibliometric studies.

The rest of this paper is organized as follows: related works are summarized in a Literature Review, the section 3 presents the SAO identification method followed by an empirical study on SCIM. Finally, we conclude our study and address future work.

2. Literature review

There are usually two kinds of SAO extraction approaches: the symbolic approach and the statistical approach.

2.1. Symbolic SAO extraction approach

A symbolic approach consists of a set of rules, often hand-written but sometimes automatically learned, that model different language phenomena. This approach focuses on the designing of rules and is popular and efficient. The Knowledgist™2.5 (or GoldFire) is probably the most well-known rule-based SAO extraction tool and has been used in many SAO analyses.[5, 6, 22, 23] However, Knowledgist™2.5 (or Goldfire) is more of a retrieval tool for patent analysts or technical developers, and cannot be easily used in quantitative analysis. It does not support mass SAO extraction. People have to enter search terms, and only the specific SAOs that contain the search terms can be obtained. Some researchers try to design their own rules to extract specific SAOs or variations of SAO (including Resource Description Frameworks (RDF), semantic relation structure). X. Wang *et al*[24] and J. Guo *et al*[25] design a set of SAO extraction rules based on Stanford parse software. G. Cascini *et al*[4] designed a syntactic parser to perform SAO extraction, and then identified the core components of patents based on SAO. T. Jiang *et al*[26] and A. Ben *et al*[27] designed a set of syntactic patterns to identify specific RDFs.

In summary, symbolic approach is unsupervised and does not require a great quantity of manual annotation. The extraction process is fairly intuitive and controllable. One can change a rule or create a new one if, after testing, he sees that something is wrong or missing. However, a symbolic approach is relatively expensive because you need to manually design a lot of rules. The precision may experience huge fluctuations dependent upon the rules.

2.2. Statistical SAO extraction approach

A statistical approach typically uses a mathematic statistical model and machine learning algorithms to learn the language phenomena.[28, 29] M. Bundschus *et al*[30] extended the framework of Conditional Random Fields to perform the annotation and extraction of semantic relations, which is the key to SAO identification. J. Punuru, J. H. Chen[31] propose an unsupervised technique for extracting SAO from domain texts. A statistical method with log-likelihood ratios is used to estimate the significance of relationships between concepts and to select suitable relation labels. D. Gerber *et al*[32] presented a statistical method in combination with an unsupervised, as well as a supervised, machine learning technique to extract RDF (a variation of SAO) triples from unstructured data streams.

In summary, statistical approach is becoming popular in SAO extraction. This approach can find the inherent law in syntax and filter out the incorrect syntax based on the statistical model. However, there are some limitations: (1) The dependence on Named Entity Recognition.[33] It is difficult to identify all kinds of concepts (named entity) in corpora based on Named Entity Recognition, and that makes it impossible to extract all SAO. (2) The statistical approach is a black box; it is less direct and less intuitive when improving the quality of SAO extraction algorithms. (3) Supervised machine learning methods need a lot of manual annotation.

3. Methodology

This paper constructs an SAO identification method that aims to improve the quality (relationship with topic, and synonyms in SAO) of identification results. Figure 1 shows the process: (1) core components of SAO structures are identified from ST&I (Science, Technology & Innovation) records based on a co-word algorithm and term clumping; (2) SAO structures are extracted based on a syntax tree-based hierarchical model; and (3) an SAO weighting model evaluates and ranks the SAO structures for selection.

3.1. Components identification model

We introduce term clumping and design a co-word algorithm to perform SAO components identification. SAO components are a set of topic terms that are used as the candidate terms for the subject and object of SAO. There are three steps to identify SAO components: (1) performing term clumping to obtain a set of words/phrases; (2) ranking the term clumping results based on the proposed co-word algorithm; and (3) combining the top N term clumping results with keywords to obtain the core components of SAO. The detailed procedure is shown below:

(1) Term clumping

To achieve these components, we introduce term clumping. Term clumping is the series of steps to clean and consolidate rich sets of topical phrases and terms in a collection of documents relating to a topic of interest [15]. We applied such term clumping steps to our data. The stepwise actions are noted in Table 2. Y. Zhang *et al*[15] shows the detailed description of each step. Term clumping typically results in mostly terms, which are useful for identifying topic terms, but do contain a great deal of noise which doesn't have a close relationship with topics.[34]

(2) Ranking the term clumping results based on co-word algorithms

To overcome the problem (noise in term clumping results) above, a co-word algorithm is proposed to evaluate and rank the results of term clumping. If a term co-occurs with keywords frequently in many papers, but does not occur so frequently itself in papers (not a common word), then we can conclude that this term is a relatively important word/phrase. This algorithm calculates the degree of importance of terms built on the co-occurrences with keywords and their own occurrence frequency. The detailed description of this algorithm is 1) the algorithm calculates the sum of co-occurrence frequency between term t and every keyword; 2) the algorithm divides the sum of co-occurrence frequency above by the number of instances of term t in papers to produce a weight. This weight is used to evaluate and rank the terms in term clumping results. The algorithm is implemented in Java and GATE, and described as below:

$$W_t = \sum_{k \in S_k} \frac{freq(t,k)}{I_t} \quad (1)$$

t: Term t derived from term clumping, $t \neq k$, $t \in S_t$.

W_t: The weight of term t used to evaluate the importance of term t.

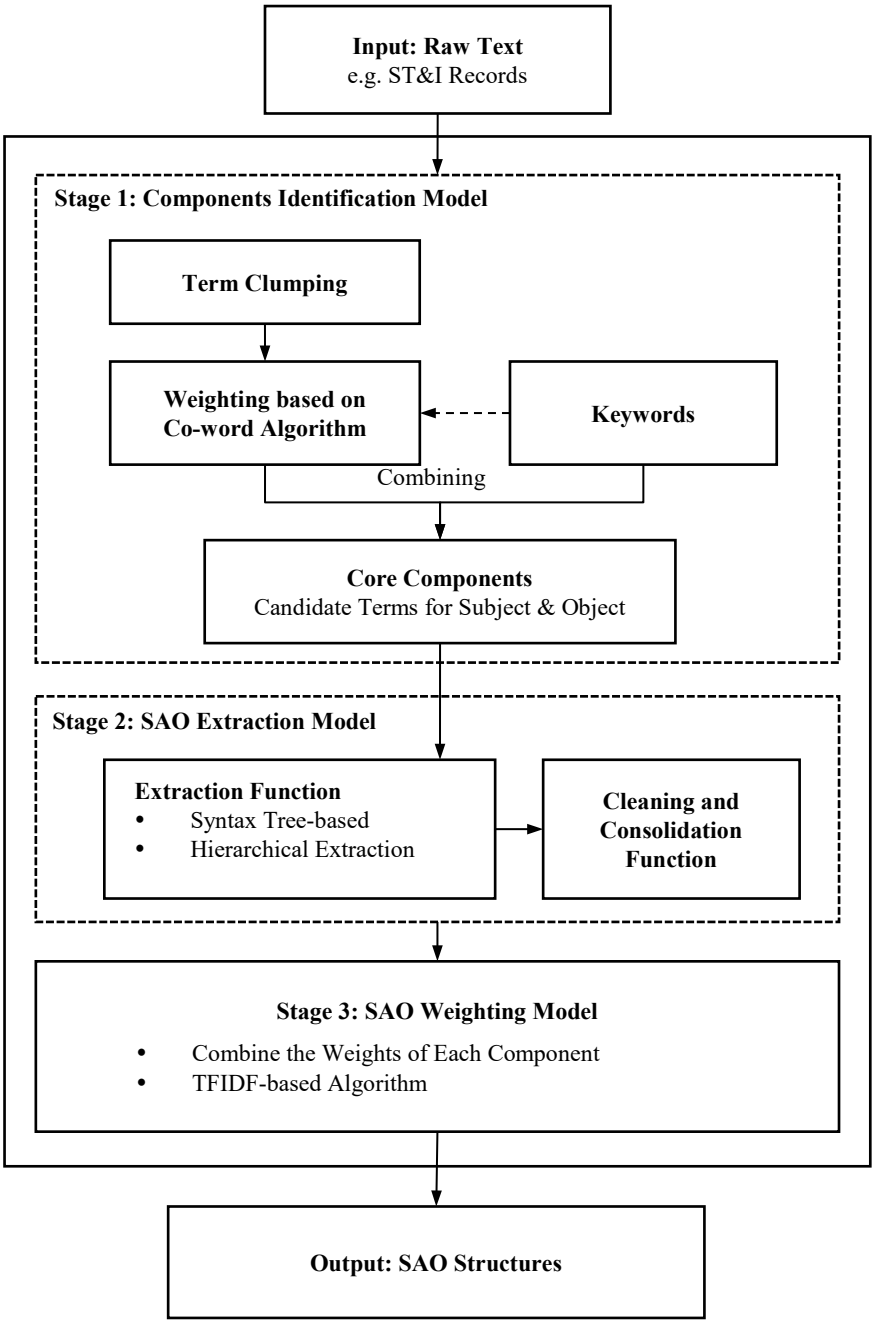

Fig. 1. Framework of extracting and ranking SAO

I_t: Total number of instances of term t in the dataset.

k: Keyword k, $k \in S_k$.

S_k: The set of keywords.

$\mathrm{freq}(t, k)$: Frequency of co-occurrence between term t and keyword k.

The benefit of ranking term clumping results based on the proposed co-word algorithms is that we consider the importance of the terms based on the relevance of terms to keywords. Keywords can show the topic of text[35, 36] and is good at forming SAO structures. The terms co-occurring with keywords frequently (rank

highly) usually can be combined with keywords to form a SAO. Compared with TFIDF, the proposed method focus on the co-occurrence of terms and keywords instead of only the occurrence of terms/records, and is helpful for construct SAO structure based on keywords.

(3) Combining the top (higher weight) term clumping results with keywords

The keywords in papers are often accurate and clean. In order to cover all the SAOs, the proposed method combines the term clumping results with the keywords.

The terms in term clumping results S_t is ranked based on W_t, and then are combined with keywords to produce the SAO core components S_c:

$$S_c = S_k + \{t | t \in \text{top } x \text{ of } S_t \text{ ranking based on } W_t\} \quad (2)$$

S_t: the set of term clumping results.

S_c: the set of SAO core components.

At last, we will check the results and remove the common terms manually.

3.2. SAO extraction model

An SAO extraction model is constructed, which is hierarchical and syntax-tree based (shown in Fig. 2 and Fig. 3). The SAO core components (obtained in section 3.1) are imported into this model to assure the subject/object of SAO has a close relationship with the topic. The model is hierarchical and equipped with specialized syntax rules (shown in Table 1) to meet the special extraction requirements which is some different with general SAO extraction (e.g., The desired SAO of "robotics technology holds a significant promise for improving industrial automation" is "robotics technology improves industrial automation" not "robotics technology holds a significant promise"). The proposed hierarchical model solves the extraction work by breaking it down into a collection of different levels of simpler sub works, like the identification of subject, action, and object. The SAO extraction model is implemented in GATE [37] and comprises six steps (shown in Figure 3, and corresponding to a different pseudo code in Figure 2):

(1) To reach the desired SAO, we first identify the objects in the sentence (including clauses) based on the syntax tree and the core components acquired in Stage 1. This step corresponds to 1-5 in Figure 2.

(2) The action is identified using the object from Step 1. Because actions and objects are built naturally upon each other, a set of rules filters actions and objects in a recursive manner. This step corresponds to 6-8 in Figure 2.

(3) Objects and actions are combined to produce a reasonable action-object structure. This step corresponds to 9-14 in Figure 2.

(4) Based on the action of the action-object structure (obtained above) and core components, we retrieve the subject of the various level of topic information. This step corresponds to 15 and 16 in Figure 2.

(5) The subject is combined with the action-object structure to produce a complete and accurate Subject-Action-Object. This step corresponds to 17-23 in Figure 2. We use the stem of verb and noun to form SAO structures to improve the "synonyms in SAO".

(6) The SAO is cleaned and consolidated using thesaurus and fuzzy matching. First, a stop word list is used to remove common SAOs. Secondly, a thesaurus of synonyms (including verbs and nouns) is constructed to combine similar SAO components (Subjects, Actions and Objects). Finally, we use fuzzy matching to combine similar SAOs. This step is fulfilled with VantagePoint.[38] It is useful to improve the "synonyms in SAO".

Algorithm: SAO Extraction

1: **For** each Seq_w **do**
2: **If** Seq_w match Rule 1 **then**
3: $TP_c \leftarrow Seq_w$
4: **If** verb + TP_c match Rule 2 and $TP_c \in S_c$ **then**
5: $Obj_c \leftarrow TP_c$
6: **For** each TP_c **do**
7: **If** verb + TP_c match Rule 3 **then**
8: $Act_c \leftarrow verb$
9: **For** each Act_c **do**
10: **If** $Act_c + TP_c$ match Rule 4 **then**
11: **If** $TP_c = Obj_c$ **then**
12: $Act_{cf} \leftarrow Act_c$ and
13: $Obj_{cf} \leftarrow Obj_c$ and
14: $AO_c \leftarrow Act_{cf} + Obj_{cf}$
15: **If** $TP_c + Act_c$ match Rule 5 and $TP_c \in S_c$ **then**
16: $Sub_c \leftarrow TP_c$
17: **For** each AO_c **do**
18: **If** $Sub_c + AO_c$ match Rule 6 **then**
19: SAO $\leftarrow Sub_c + AO_c$ and
20: Sub $\leftarrow Sub_c$ and
21: Act $\leftarrow Act_{cf}$ and
22: Obj $\leftarrow Obj_{cf}$
23: **Else return** non-complete SAO

Fig. 2. The pseudo code for SAO extraction Model

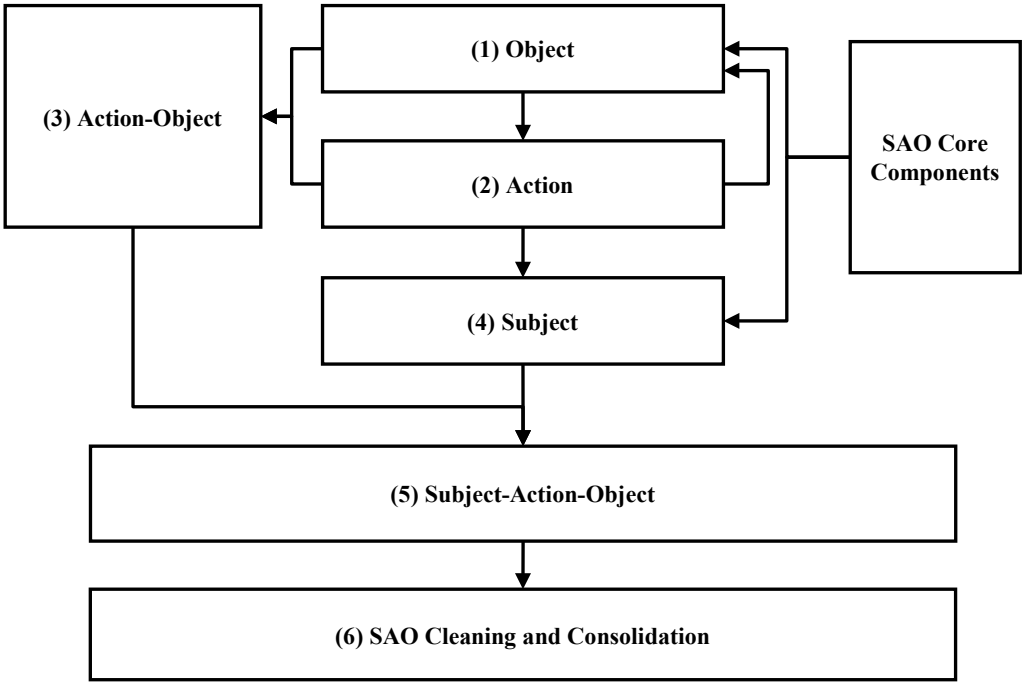

Fig. 3. SAO extraction Model

In Figure 2, COR denotes the corpus that contains all the records; Seq_w denotes words sequence ∈ COR; TP_c denotes candidate word/phrase; Obj_c denotes candidate Object; Obj_{cf} denotes further candidate Object got by filtering Obj_c ; Act_c denotes candidate Action; Act_{cf} denotes further candidate Action thatis produced by filtering Act_c; AO_c denotes candidate Action + Object; Sub_c denotes candidate Subject; SAO denotes Subject + Action + Object; Sub denotes the final Subject; Act denotes the final Action; Obj denotes the final Object.

The Rule 1-6 embedded in the pseudo code in Figure 2 are a series of judgment conditions (syntax rules) which are used to judge different levels of SAO structure, (e.g., if a string of words fit the pattern "$(JJ|VBN)[0,3] + (N \cap N.length > 1)[1,m]$" in Rule 1, then it is identified as candidate word/phrase (TP_c); if a TP_c satisfy Rule 2, then it is identified as candidate object of SAO (Obj_c)). Rules 1-6 are syntax tree-based and written in Jape (shown in Table 1).

Table 1. Rules mentioned in pseudo code of SAO extraction Model

Rule	Description
Rule 1	Identify TP_c, e.g. $(JJ\|VBN)[0,3] + (N \cap N.length > 1)[1,m]$ is a pattern to match TP_c.
Rule 2	Identify Obj_c, e.g. V + (token ∩ token ≠ N ∩ token ≠ punctuation)[0,5] + TP_c, if this pattern is matched, TP_c is identified as Obj_c. Rule 2 includes 32 patterns.
Rule 3	Identify Act_c, e.g. $(VBZ\|VBP\|VB\|VBD\|VBG) \cap (V \neq$ "be") + (token ∩ token ≠ V)[0,5] + TP_c, if this pattern is matched, The verb (VBZ|VBP|VB|VBD|VBG) in this pattern is Act_c. Rule 3 includes 7 patterns.
Rule 4	Identify AO_c, e.g. $Act_c + (tokens \cap tokens \neq Act_c \cap tokens \neq Obj_c)[0,m] + Obj_c^1 + Obj_c^2$, if this pattern is matched, there are two AO_c: $Act_c + Obj_c^1$ and $Act_c + Obj_c^2$.
Rule 5	Identify Sub_c , e.g. $TP_c + (token =$ "as")? $+MD? + R? + V$, if this pattern is matched, TP_c is equivalent to Sub_c. Rule 5 include 22 patterns.
Rule 6	Identify final SAO, e.g. $Sub_c^1 + (token =$ "and") $+ Sub_c^2 + (tokens \cap tokens \neq Sub_c \cap tokens \neq AO_c)[0,m] + AO_c$, if this pattern is matched, there are two SAO: $Sub_c^1 + AO_c$ and $Sub_c^2 + AO_c$.

In Table 1, JJ indicates the adjective. VBN indicates the past participle. N indicates the noun. V indicates the verb. VBZ indicates the 3rd person singular present. VBP indicates the non-3rd person singular present. VB indicates the base form of the verb. VBD indicates the past tense of the verb. VBG indicates the gerund or present participle. MD indicates the modal. R indicates the adverb.

3.3. SAO weighting model

An SAO weighting model is constructed to evaluate the importance of each SAO obtained in section 3.2. This model ranks and identifies important SAOs for subsequent analysis (e.g., topic analysis).

The SAO weighting model is constructed based on the idea of TFIDF. TFIDF reflects the important a term to a document in a corpus based on the frequency of the term and records.[39] SAO is a combination of terms and verbs, so we can combine the TFIDF of each component to calculate the weight of SAO. In other words, to calculate the SAO weight, the weighting model calculates the TFIDF of each part (subject, object, action, and SAO as a whole) and combines them together. Subject and Object are the core of SAO and in a similar position. Action can link Subject and Object together, and has great potential to affect the overall weight. Thus, we add the TFIDF of Subject and Object together, and use the weight of Action as a global parameter. Meanwhile, there is another global parameter. We call it the initial weight of SAO, where we treat SAO as a term and calculate the TFIDF of overall SAO (IW_{SAO}). Finally, the weight of the SAO is calculated as follows:

$$W_{SAO} = (W_S + W_O) * W_A * IW_{SAO} \quad (3)$$

in which W_{SAO} is the weight of the SAO, W_S is the weight of the Subject, W_O is the weight of the Object, W_A is the weight of the Action, IW_{SAO} is the initial weight of the SAO. W_S, W_O and IW_{SAO} are calculated based on TFIDF:

$$W_S = \log(1 + I_S) * \log\left(1 + \frac{N}{R_S}\right) \quad (4)$$

$$W_O = \log(1 + I_O) * \log\left(1 + \frac{N}{R_O}\right) \quad (5)$$

$$IW_{SAO} = \log(1 + I_{SAO}) * \log\left(1 + \frac{N}{R_{SAO}}\right) \quad (6)$$

In these formulas, I_S denotes instances, or the total number of times the subject appears in the dataset; R_S denotes the number of records that contain this subject; I_O denotes the total number of instances the object appears in the dataset; R_O denotes the number of records that contain that Object; I_{SAO} denotes the total number of instances the SAO appears in the dataset; R_{SAO} denotes the number of records that contain that SAO; N indicates the total number of documents in the dataset. The calculations of W_S, W_O and IW_{SAO} are based on TFIDF. Take W_S for example, $\log(1 + I_S)$ is the logarithmically scaled Subject frequency, and $\log\left(1 + \frac{N}{R_S}\right)$ is the logarithmically scaled inverse document frequency. Consequently, if a Subject appears frequently in the whole data set but only a small number of records contain this Subject, then this Subject has a higher weight.

For the weight of the Action, we find it is different with Subject and Object. The term frequency and document frequency of the Action cannot express its importance. Thus, we first identify important Actions based on the statistics of verbs. Then W_A is set via expert knowledge.

4. A case study

We applied the proposed SAO identification method to the publications in the *Journal of Scientometrics*. *Scientometrics* is a leading journal in the field of Information Science & Library Science, and provides good balances between theoretical research & empirical studies, and information science & management needs. Such publications, which contain a rich variety of topics with strong features of coupling, would make great sense to be used for our SAO identification and explore insights to compare with the traditional bibliometric techniques (word-based). In light of this, we introduce our method to identify the SAO. In total, 4,215 *Scientometrics* papers authored between 1978 and 2015 were collected from Web of Science (Search query: SO=Scientometrics, Query Date: 13 08 2015).

4.1. Components identification

This section displays the process of how to obtain the core components of SAO. There are three steps, as described in section 3.1.

The results from the first step—term clumping—are listed in Table 2. These form the basis for core component identification.

Table 2. Term clumping stepwise results

Field selection	*Number of phrases and terms*
Phrases with which we begin	57,205
Basic Term Cleaning—remove common terms	53,888
Basic Term Cleaning—remove general scientific terms	52,302
Term Consolidation—fuzzy matching	44,542
Association Rule-based Consolidation—combine low frequency terms with high frequency terms that frequently appear in the same record	31,952
Association Rule-based Consolidation—combine terms that share 3 or more words	27,263
Term Consolidation—fuzzy matching	26,192
Pruning—remove terms that appear only in one record	7,569

In the second step, we applied co-word algorithm mentioned in section 3.1 to rank the results of term clumping. The top 10 is shown as sample in Table 3. The weight W_t of term clumping result terms in Table 3 has been normalized.

Table 3. Weighting of term clumping results

Number	Term clumping result term	Weight
1	hub/authority scores	0.916667
2	high impact articles	0.861111
3	document co-citation network	0.833333
4	environmental engineering	0.805556
5	regional innovation system	0.75
6	international collaboration output	0.666667
7	co-word network	0.638888889
8	country publication share	0.611111111
9	food science	0.611111111
10	economic cooperation	0.583333333

Finally, by combing the top 80% of term clumping results (6,055 words/phrases) with keywords set (4,003 words/phrases), we achieved the core components (8,762 words/phrases). There is overlap between term clumping results and keywords, so the number of core components is not the sum of them. We chose the top 80% based on previous experience and experts' knowledge. With this level, we will not lose the important words/phrases and still filter out the common ones that do not have a close relationship with the topic of interest.

4.2. SAO extraction and SAO weighting

From these core components we obtained 89,713 SAO structures based on the SAO extraction model. These SAOs were cleaned and consolidated via fuzzy matching to produce a final result of 84,241 SAOs. Based on the statistics of verbs and expert knowledge, we obtained 169 core Action words in graphene field, and then W_A is set via expert knowledge. Every SAO was evaluated by the SAO weighting model and the top 10 results are shown in Table 4. Weights have been normalized.

Precision and recall is introduced to validate the results and test the reliability. Precision measures the number of correctly identified SAOs as a percentage of the number of SAOs identified. The higher the precision, the better the system is at ensuring that what is identified is correct. Recall measures the number of correctly identified SAOs as a percentage of the total number of correct SAOs. The higher the recall rate, the better the system is at not missing correct SAOs. A random 100 papers was selected to form the test dataset. We annotated the test dataset manually to identify the SAO structures and use it as 'gold standard' against which to compare the proposed SAO identification method. Based on the 'gold standard', the precision and recall of SAOs in each paper are calculated, and then we calculated the average of the precision and recall. The average of precision is 0.8058, the average of recall is 0.8446.

Table 4. Top 10 SAO (weighted)

SAO	Subject	Action	Object	SAO weight
Factors affect research productivity	Factors	affect	research productivity	1
Author co-citation analysis discover intellectual structure	Author co-citation analysis	discover	intellectual structure	0.996136316
Citations contribute impact factor	Citations	contribute	impact factor	0.975317836
Thomson Reuters compute JIF	Thomson Reuters	compute	JIF	0.939696411
Methodology identify Frontier Areas	methodology	identify	Frontier Areas	0.89139095
Academic research focus China	academic research	focus	China	0.881187012
Factors explain collaboration	Factors	explain	collaboration	0.84504439
Patents result in international collaboration	Patents	result in	international collaboration	0.828255483
Aim map intellectual structure	Aim	map	intellectual structure	0.789898655
cluster analysis reveal scientists	cluster analysis	reveal	scientists	0.782380707

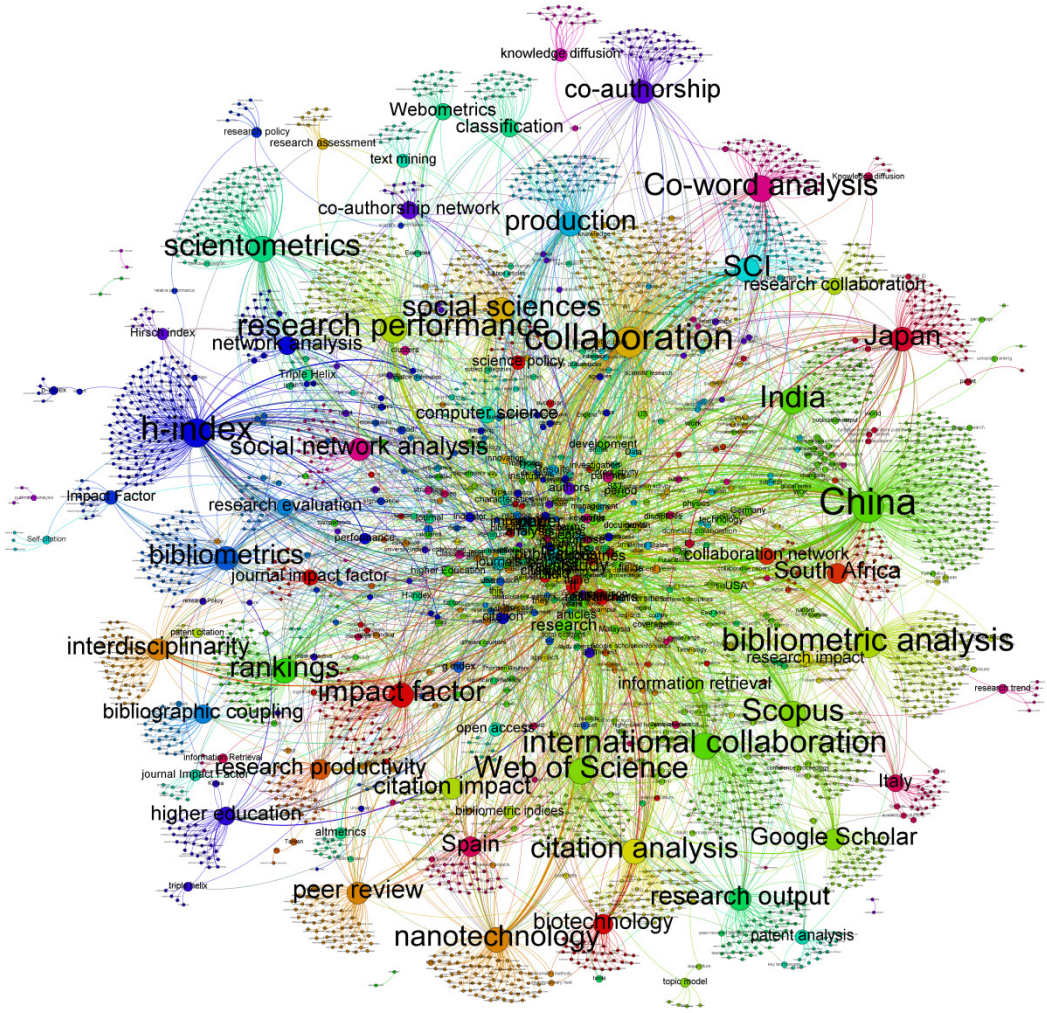

Fig. 4. Result of SAO identification

A network of the SAOs is constructed to show the SAO identification results. The nodes in the network represent Subjects or Objects, and the edges represent Actions between Subjects and Objects (shown as Fig. 4). The size of the node reflects the frequency of the term in the SAO set. The strength of the lines is related to the number of SAOs that contain two nodes together. Given the sheer quantity of data, Fig. 4 displays the SAOs that have a relation with the top 70 keywords.

Based on the in-degree and out-degree of nodes, we find that "collaboration" is the most popular topic in *Journal of Scientometrics*. Other hot topics include co-word analysis, co-authorship analysis, network analysis, h-index, citation analysis, and impact factor. The link between nodes indicates the real relationship between topics. For example, "Co-word analysis map collaboration" is an SAO and indicates that "Co-word analysis" method can be used to "map" the "collaboration" situation.

The proposed method has three advantages:

(1) The SAOs identified with the proposed method contains a wealth of semantic information and indicates a specific relationship between topic terms; therefore, each topic is described in detail by a set of weighted SAOs, which show its core content.

(2) The SAOs identified with the proposed method have a close relationship with the topics, which makes it very useful for topic analysis.

(3) The SAOs identified in the proposed method improve the "synonyms in SAO", and can be used in quantitative analysis, including: 1) the statistics of the SAO frequency, Subject frequency, Action frequency, Object frequency; 2) building the SAO-SAO matrix, SAO-document matrix, SAO-author matrix, SAO-coutry matrix, SAO-year matrix and SAO-organization matrix, and then implementing co-occurrence analysis; and 3) building the Subject-Action-Object network and implementing SAOs' network analysis.

5. Discussion and conclusions

Existing SAO research have predominantly focused on applications rather than the SAO extraction techniques. The limitation is that SAO structures cannot fit the application in bibliometric analysis. This paper proposes an SAO identification approach that combines a symbolic and statistical approach.

We find there are two challenges in SAO identification for bibliometrics: low relevance of SAOs to domain topics; and synonyms in SAO. Aiming to solve these two problems, the proposed method introduces a component identification model, a hierarchical SAO extraction model, and an SAO weighting model. With the help of three models, researchers can identify complete and accurate SAO structures. At the same time, The SAOs identified with the proposed method contain a wealth of semantic information and can indicate the relationship between topic terms.

Compared with existing SAO identification methods, the proposed method aims to provide basis for SAO-based bibliometric analysis, focus on the components (Subject, Action, Object) of SAO and the topic relevance. We emphasize on the quantitative SAO analysis.

There are many possible applications for our method: (1) identify topic terms and rank topics in order of popularity and influence; (2) classify topics based on the similarity of relationship; (3) provide in-depth functional descriptions of topics via corresponding SAOs; and (4) explore the relationships between topics based on the action between topic terms.

However, there are also some limitations to this paper. We use traditional fuzzy matching to combine similar SAOs, which is not good at processing long and complex SAO structures. It is better to introduce ontology for combining similar SAOs. We engaged experts for setting indictor thresholds, but a systematic setting process would be able to improve the efficiency of qualitative approaches.

In future research, we will further apply our approaches to various topics and texts to test its robustness. In addition, we will continue to improve the cleaning and consolidation function for combining similar SAOs without affecting their original semantic information. We also see potential in associating SAO-based network analysis with technology roadmapping to analyze topic development trends.

Acknowledgements

We acknowledge support from the General Program of National Natural Science Foundation of China (Grant No. 71373019). This paper was also funded by the International Graduate Exchange Program of Beijing

Institute of Technology. The findings and observations contained in this paper are those of the authors and do not necessarily reflect the views of the supporters.

References

1. S. Auer, J. Lehmann, Creating knowledge out of interlinked data, *Semantic Web.* **1**(1) (2010) 97-104.
2. R. C. Gudivada, X. Y. A. Qu, J. Chen, A. G. Jegga, E. K. Neumann, B. J. Aronow, Identifying disease-causal genes using Semantic Web-based representation of integrated genomic and phenomic knowledge, *Journal of Biomedical Informatics.* **41**(5) (2008) 717-729.
3. Y. Zhao, S. Gao, P. Gallinari, J. Guo, Knowledge base completion by learning pairwise-interaction differentiated embeddings, *Data Mining and Knowledge Discovery.* **29**(5) (2015) 1486-1504.
4. G. Cascini, A. Fantechi, E. Spinicci, Natural language processing of patents and technical documentation, In: Marinai S, Dengel A, eds. *Document Analysis Systems VI.* Vol 3163. (Springer Berlin Heidelberg, Berlin, 2004), 508-520.
5. M. G. Moehrle, L. Walter, A. Geritz, S. Muller, Patent-based inventor profiles as a basis for human resource decisions in research and development, *R & D Management.* **35**(5) (2005) 513-524.
6. S. Choi, J. Yoon, K. Kim, J. Y. Lee, C. H. Kim, SAO network analysis of patents for technology trends identification: a case study of polymer electrolyte membrane technology in proton exchange membrane fuel cells, *Scientometrics.* **88**(3) (2011) 863-883.
7. Y. Kim, Y. Tian, Y. Jeong, R. Jihee, S.-H. Myaeng, Automatic discovery of technology trends from patent text. *2009 ACM Symposium on Applied Computing*. (ACM, Honolulu, Hawaii, 2009), 1480-1487.
8. Y. Zhang, X. Zhou, A. L. Porter, J. M. V. Gomila, A. Yan, Triple Helix innovation in China's dye-sensitized solar cell industry: hybrid methods with semantic TRIZ and technology roadmapping, *Scientometrics.* **99**(1) (2014) 55-75.
9. L. Leydesdorff, L. Vaughan, Co‐occurrence matrices and their applications in information science: extending ACA to the web environment, *Journal of the American Society for Information Science and technology.* **57**(12) (2006) 1616-1628.
10. S. Ravikumar, A. Agrahari, S. Singh, Mapping the intellectual structure of scientometrics: a co-word analysis of the journal Scientometrics (2005–2010), *Scientometrics.* **102**(1) (2015) 929-955.
11. H. N. Su, P.-C. Lee, Mapping knowledge structure by keyword co-occurrence: a first look at journal papers in Technology Foresight, *Scientometrics.* **85**(1) (2010) 65-79.
12. P. Van Den Besselaar, G. Heimeriks, Mapping research topics using word-reference co-occurrences: A method and an exploratory case study, *Scientometrics.* **68**(3) (2006) 377-393.
13. D. J. d. Price. Little science, big science: New York: Columbia University Press; 1963.
14. Y. Zhang, L. Shang, L. Huang, et al., A hybrid similarity measure method for patent portfolio analysis, *Journal of Informetrics.* **10**(4) (2016) 1108-1130.
15. Y. Zhang, A. L. Porter, Z. Hu, Y. Guo, N. C. Newman, "Term clumping" for technical intelligence: a case study on dye-sensitized solar cells, *Technological Forecasting and Social Change.* **85** (2014) 26-39.
16. Y. Ding, G. G. Chowdhury, S. Foo, Incorporating the results of co-word analyses to increase search variety for information retrieval, *Journal of Information Science.* **26**(6) (2000) 429-451.
17. H. Peters, A. F. van Raan, Co-word-based science maps of chemical engineering. Part I: Representations by direct multidimensional scaling, *Research Policy.* **22**(1) (1993) 23-45.
18. L. Vaughan, J. You, Word co-occurrences on Webpages as a measure of the relatedness of organizations: A new Webometrics concept, *Journal of Informetrics.* **4**(4) (2010) 483-491.
19. H. J. Peat, P. Willett, The limitations of term co-occurrence data for query expansion in document retrieval systems, *JASIS.* **42**(5) (1991) 378-383.
20. Y. Zhang, G. Zhang, H. Chen, A. L. Porter, D. Zhu, J. Lu, Topic analysis and forecasting for science, technology and innovation: Methodology with a case study focusing on big data research, *Technological Forecasting and Social Change.* **105** (2016) 179-191.
21. H. Chen, G. Zhang, D. Zhu, J. Lu, A patent time series processing component for technology intelligence by trend identification functionality, *Neural Computing and Applications.* **26**(2) (2015) 345-353.
22. I. Bergmann, D. Butzke, L. Walter, J. P. Fuerste, M. G. Moehrle, V. A. Erdmann, Evaluating the risk of patent infringement by means of semantic patent analysis: the case of DNA chips, *R&D Management.* **38**(5) (2008) 550-562.
23. S. Choi, H. Kim, J. Yoon, K. Kim, J. Y. Lee, An SAO-based text-mining approach for technology roadmapping using patent information, *R & D Management.* **43**(1) (2013) 52-74.
24. X. Wang, P. Qiu, D. Zhu, L. Mitkova, M. Lei, A. L. Porter, Identification of technology development trends based on subject–action–object analysis: The case of dye-sensitized solar cells, *Technological Forecasting and Social Change.* **98** (2015) 24-46.
25. J. Guo, X. Wang, Q. Li, D. Zhu, Subject–action–object-based morphology analysis for determining the direction of technological change, *Technological Forecasting and Social Change.* **105** (2016) 27-40.
26. T. Jiang, A. H. Tan, K. Wang, Mining generalized associations of semantic relations from textual Web content, *Ieee Transactions on Knowledge and Data Engineering.* **19**(2) (2007) 164-179.

27. A. Ben Abacha, P. Zweigenbaum, Automatic extraction of semantic relations between medical entities: a rule based approach, *Journal of biomedical semantics.* **2 Suppl 5** (2011) S4.
28. H. Li, X. Wu, Z. Li, G. Wu, A relation extraction method of Chinese named entities based on location and semantic features, *Applied Intelligence.* **38**(1) (2013) 1-15.
29. L. Lu, L. I. Bi-Cheng, Named entity relation extraction based on SVM training by positive and negative cases, *Journal of Computer Applications.* **28**(6) (2008) 1444-1437.
30. M. Bundschus, M. Dejori, M. Stetter, V. Tresp, H. P. Kriegel, Extraction of semantic biomedical relations from text using conditional random fields, *Bmc Bioinformatics.* **9**(1) (2008) 1-14.
31. J. Punuru, J. H. Chen, Learning non-taxonomical semantic relations from domain texts, *Journal of Intelligent Information Systems.* **38**(1) (2012) 191-207.
32. D. Gerber, S. Hellmann, L. Buhmann, T. Soru, R. Usbeck, A. C. N. Ngomo, Real-Time RDF Extraction from Unstructured Data Streams, In: Alani H, Kagal L, Fokoue A, et al., eds. *Semantic Web - Iswc 2013, Part I.* Vol 8218. (Springer-Verlag Berlin, Berlin, 2013), 135-150.
33. T. H. Cao, NAMED ENTITY DISAMBIGUATION: A HYBRID APPROACH, *International Journal of Computational Intelligence Systems.* **5**(6) (2012) 1052-1067.
34. H. Chen, Y. Zhang, G. Zhang, D. Zhu, Modeling technological topic changes in patent claims, in *Portland International Conference on Management of Engineering and Technology.* (IEEE, Portland, 2015), 2049-2059.
35. H. Chen, G. Zhang, J. Lu, D. Zhu, A fuzzy approach for measuring development of topics in patents using Latent Dirichlet Allocation. *IEEE International Conference on Fuzzy Systems*. (IEEE, Turkey, 2015), 1-7.
36. H. Chen, Y. Zhang, D. Zhu, Identifying Technological Topic Changes in Patent Claims Using Topic Modeling, In: Daim TU, Chiavetta D, Porter AL, Saritas O, eds. *Anticipating Future Innovation Pathways Through Large Data Analysis*. (Springer International Publishing, Cham, 2016), 187-209.
37. H. Cunningham, V. Tablan, A. Roberts, K. Bontcheva, Getting More Out of Biomedical Documents with GATE's Full Lifecycle Open Source Text Analytics, *PLoS Comput Biol.* **9**(2) (2013) e1002854.
38. VantagePoint. Available at: www.theVantagePoint.com. Accessed 19 November, 2016.
39. H. Noh, Y. Jo, S. Lee, Keyword selection and processing strategy for applying text mining to patent analysis, *Expert Systems with Applications.* **42**(9) (2015) 4348-4360.