Automatic Topic Identification for Idea Summarization in Idea Visualization Programs

Kobkrit VIRIYAYUDHAKORN\(^{(a)}\), Nonmember and Susumu KUNIFUJI\(^{(b)}\), Member

SUMMARY Recent idea visualization programs still lack automatic idea summarization capabilities. This paper presents a knowledge-based method for automatically providing a short piece of English text about a topic to each idea group in idea charts. This automatic topic identification makes use of Yet Another General Ontology (YAGO) and Wordnet as its knowledge bases. We propose a novel topic selection method and we compared its performance with three existing methods using two experimental datasets constructed using two idea visualization programs, i.e., the KJ Method (Kawakita Jiro Method) and mind-mapping programs. Our proposed topic identification method outperformed the baseline method in terms of both performance and consistency.

key words: creativity support system, idea summarization, KJ method, knowledge-based automatic topic identification, mind map

1. Introduction

Creativity support systems are computer systems that involve the storage, analysis, and representation of ideas in a form that facilitates human interaction to aid creativity, and they are increasingly important for competitive creative economies. An idea visualization program is a type of creativity support system that represents ideas using a tree structure, which facilitates hierarchical and categorical idea organization. These idea visualization programs were developed based on popular creativity techniques, such as the KJ Method (Kawakita Jiro Method)\(^{(1)}\), and mind mapping.

Automatic idea summarization is a computer support of idea processing in the generative support level, which is the highest computer support level for idea processing system firstly coined by L.F. Young in 1988\[^{2}\]. Current natural language processing technology facilitates automatic idea summarization in creativity support systems, which may be implemented in practice with acceptable accuracy. Automatic topic identification, which is a method for identifying the central ideas in a text, is used to achieve this task. Automatic topic identification can be classified into three categories: statistical approaches, knowledge-based approaches, and hybrid approaches. The amount of text that should be produced by idea visualization programs is typically short. Therefore, the external knowledge is needed to identify relevant topics. Thus, the proposed method is a knowledge-based approach, which uses two general domain knowledge bases, i.e., Wordnet\[^{3}\] and Yet Another General Ontology (YAGO)\[^{4}\].

In this paper, we propose a novel topic selection method for knowledge-based automatic topic identification, which is optimized to identify topics for idea groups in idea visualization programs. Our proposed method is based on a knowledge-based concept counting paradigm from a hierarchical concept taxonomy inspired by Lin’s work\[^{5}\]. The performance of our proposed topic selection method was compared with three existing proposed methods on two idea visualization programs, i.e., the KJ Method and mind-mapping programs. This paper is organized as follows. Section 2 describes related studies. Section 3 explains the idea charts and knowledge bases. Section 4 describes how our method automatically identifies topics. Section 5 presents our empirical evaluations. Section 6 is the discussion and Sect. 7 concludes the paper.

2. Related Works

Knowledge-based automatic topic identification based on concept counting paradigm from a hierarchical taxonomy has been proposed by several researchers.

Lin\[^{5}\] proposed a concept counting technique that improved an earlier term frequency counting method. Lin also proposed a topic selection method, i.e., the branch ratio, which was focused on degree of generalization of concept, and provided an efficient way to extract multiple topics from a document. Lin used Wordnet taxonomy, which contains very little knowledge about proper nouns compared with the YAGO ontology. Tiun et al.\[^{6}\] proposed a topic selection method, i.e., the ratio balance, which ensure that the accumulated weight was higher than that expected by using the Yahoo ontology as the knowledge base.

Automatic topic identification based on information retrieval techniques has been proposed by many authors. He et al.\[^{7}\] proposed using an unsupervised clustering technique. Özmutlu et al.\[^{8}\] proposed the use of a neural network classification technique. Clifton et al.\[^{9}\] proposed the use of an association rule mining technique. Coursey\[^{10}\] proposed the use of an improved Pagerank algorithm for topic identification with Wikipedia. Muñoz-García et al.\[^{11}\] proposed the use of semantic relatedness and disambiguation techniques based on DBpedia.

Research studies based on artificial intelligence tools for idea visualization programs have been proposed by several researchers. Leake et al.\[^{12}\] and Kobkrit et al.\[^{13}\]...
proposed knowledge-based methods that suggested relevant information to assist the production of idea charts. Vol-
lalon [14] proposed an automatic concept map generation
based on short essays.

3. Idea Chart and Knowledge Bases

3.1 Structure of Idea Chart

An idea chart is a representation of a set of ideas that is typ-
ically drawn after performing a creativity technique for an
individual memorandum or mutual understanding in a team.
An idea chart is defined as a rooted tree as follows.

Let $T = (V, E)$ be a tree. A vertex $v \in V$ is a non-
empty word sequence $\psi \in W^*$, where $W$ is a set of words
that represents an idea. A $v$ is an internal vertex if its degree
is greater than one; otherwise, $v$ is a leaf. Let $L_T$ be the set
of leaves of $T$. An edge $e \in E$ is a hierarchical generalization
relationship; a vertex at a higher level is the generalization
of a vertex at a lower level.

The tree $T$ has exactly one root vertex that is denoted as
$R_T$, which is the title of an idea chart. The set $L_T$ is the set of
idea labels. The subset of vertices $g \subseteq V$ of an idea chart $T$,
which shares the same immediate parent vertex $h \in V - L_T$,
is a member of an idea group $G_h$. The label $h$ is a group
header of the idea group $G_h$. The set of all group headers $H$
where $h \in H$ has the following properties,

- $H = V - L_T$.
- $\cup_{h \in H}(G_h) = V - R_T$.

The structure of a sample idea chart and its tree represen-
tation is illustrated in Fig. 1. White, nonwhite, and black
labels/vertices are $L_T$, $H$, and $R_T$, respectively, where
$H = \{1, 6, 7\}$, $G_1 = \{6, 5\}$, $G_6 = \{4, 7\}$, and $G_7 = \{2, 3\}$; a number
represents a label of the sample idea chart in Fig. 1.

The research goal is to automatically identify an unla-
abeled group header $h$ in an idea group $G_h$, such as the group
header 7 (the content of which is “??”) in Fig. 1 (a). Be-
cause of the tree structure of an idea chart, the group header
$h$ should only describe the main idea of all members in $G_h$
and it should also not be more abstract than its parent node
and its siblings.

**Definition 1.** Given a group header $h$ in an idea chart $T$, the
function $\text{in}(h)$ returns a set of idea labels and their contents
are summarized in $h$, which is defined as follows.

$$\text{in}(h) = G_h$$

**Definition 2.** Given a group header $h$ in an idea chart $T$, the
function $\text{out}(h)$ returns a set of idea labels and their contents
should not be mentioned in $h$, and they are used to limit the
abstractness of $h$, which is defined as follows:

$$\text{out}(h) = G_{\text{parent}(h)} \cup \text{parent}(h) - h$$

where $\text{parent}(x)$ returns the parent of vertex $x$. For ex-
ample, if we let $h$ be the group header 7 in Fig. 1 (b): $\text{in}(h) = \{2, 3\}$ and $\text{out}(h) = \{4, 6\}$.

3.2 Knowledge Bases

Wordnet is an English lexical database that groups the
senses of words into sets of synonyms known as “synsets”
and it provides their semantic relationships, such as syn-
onymy and hypernymy [3]. YAGO is a general-purposed
semantic database with more than 10 million entities (in-
cluding common nouns, proper nouns, and name entities),
which are derived from Wikipedia, Wordnet and Geoname
ontologies. YAGO also provide the semantic relationships
among them [4]. We used Wordnet and YAGO as the exter-
nal knowledge bases for verbs and nouns, respectively.

Wordnet and YAGO are organized as ontologies. An
ontology consists of the concepts that describe things and
facts, which are the relationships between two concepts. A
fact is represented as an ordered triplet that represents a do-
main, a range, and a relationship that associates a domain
and a range. A triplet is represented as follows [4]:

$$\text{(Domain, Relations, Range)}$$

For example, to make a statement “An apple is a types of
consumable fruit.”, its triplet is defined as

$$\text{(Apple, Hyeronym, Consumable Fruit)}$$
Definition 3. Let $C$ be a finite set of concepts and $\mathcal{R}$ be a finite set of relations. A set of all possible facts $\mathcal{F}$ and an ontology $\mathcal{O}$ are defined as follows:

$$\mathcal{F} = C \times \mathcal{R} \times C$$ (3)

$$\mathcal{O} \subseteq \mathcal{F}$$ (4)

Definition 4. Given $D \subseteq C$, $c \in C$, $r \in \mathcal{R}$ and $\mathcal{O}$ is an ontology. A function $\text{fact}(c, r, \mathcal{O})$ is defined as follows.

$$\text{fact}(D, r, \mathcal{O}) = \{(d, r, x) | x \in C \land d \in D \land (d, r, x) \in \mathcal{O}\}$$ (5)

The function $\text{fact}(c, r, \mathcal{O})$ returns the set of facts that has the relation $r$ and the set of domain $D$ that appears in ontology $\mathcal{O}$.

Definition 5. Let $c \in C$ and $r \in \mathcal{R}$. Given $F \subseteq \mathcal{F}$, a function $\text{range}(F)$ is defined as follows.

$$\text{range}(F) = \{x | \exists c, r, x \in C \land (c, r, x) \in F\}$$ (6)

The function $\text{range}(F)$ returns the multiset of range concepts from the multiset of facts $F$.

4. Automatic Topic Identification

4.1 Extraction of Nouns and Verbs

Given a group header $h$ in an idea chart $T$, only the in($h$) and out($h$) sets are read as inputs. The set of nouns (including compound nouns and name entities) and a set of verbs that appear in these label sets need to be extracted. POS tagging is used to achieve this, which is the process of tagging a word with a particular lexical category (i.e., noun, pronoun, possessive ending, verb, adjective, or adverb). Only the base form of words are stored in both knowledge bases, so each word is transformed into its base form via lemmatization, e.g., children $\Rightarrow$ child, unsolved problems $\Rightarrow$ unsolved problem, and walking $\Rightarrow$ walk.

Definition 6. Let $W$ be a set of words and $\psi \in W^*$. Let $Q = \{\text{Adjective}, \text{Adverb}, \text{PossEnding}\}$ and $R = \{\text{Noun}, \text{ProperNoun}\}$ be sets of lexical categories. Let $N$ be a set of noun phrases that represents concepts in the YAGO ontology. Let $\text{pos}(X)$ return the order $n$-tuples lexical categories from text $X$. Let $\text{lemma}(X)$ returns the lemma sequence from text $X$. Given a set of labels, a function $N(L)$ is defined as follows.

$$N(L) = \bigcup_{l \in L} M(l)$$ (7)

$$M(\psi) = N \cap \{\text{lemma}(X) | X \in \text{sub}(\psi) \land \text{pos}(X) \in Q^* R^*\}$$ (8)

$$\text{sub}(\psi) = \{\psi | \exists \psi_1, \psi_2 : \psi_1' \psi_2 = \psi \land \psi' \neq \varepsilon\}$$ (9)

The multiset $N(L)$ is the multiset of all lemmatized noun phrases found in the group of label $L$. The multiset $M(\psi)$ is the multiset of lemmatized noun phrases found in a label that exist in YAGO. The set $\text{sub}(\psi)$ is the set of non-empty subsequences of $\psi$.

Definition 7. Let $W$ be a set of words and $\psi \in W^*$. Let $\mathcal{V}$ be a set of verbs in the Wordnet ontology. Given a set of labels $L$, a function $V(L)$ is defined as follows.

$$V(L) = \bigcup_{l \in L} U(l)$$ (10)

$$U(\psi) = \mathcal{V} \cap \{\text{lemma}(w) | w \in \psi \land \text{pos}(w) = (\text{verb})\}$$ (11)

the multiset $V(L)$ is the multiset of all lemmatized verbs found in the group of label $L$. The multiset $U(\psi)$ is the multiset of lemmatized verbs found in a label that exists in Wordnet.

4.2 Mapping of Words to Concepts

Next, the sets of extracted nouns and verbs are mapped to their reference concepts in YAGO and Wordnet, respectively. In YAGO, there is a relationship $\text{means}$ that maps from a noun phrase to its possible reference concepts [4], for example,

("'Apple', \text{means}, \text{AppleFruit}), ("'Apple', \text{means}, \text{AppleInc}.)\ldots"

In Wordnet, there is a relationship $\text{synsets}$ that maps from a verb to its possible concepts [3], for example,

("'Require', \text{synsets}, \text{Demand}), ("'Require', \text{synsets}, \text{Ask})\ldots"

Definition 8. By exploiting both relationships, we can extract the sets of noun and verb concepts from $N(L)$ and $V(L)$ that exist in YAGO and Wordnet respectively. Therefore, the sets of noun and verb concepts are defined as follows.

$$\text{NC}(L) = \text{range}(N(L))$$ (12)

$$\text{VC}(L) = \text{range}(V(L))$$ (13)

$$N(F) = \text{fact}(N(L), \text{means}, \text{YAGO})$$ (14)

$$V(F) = \text{fact}(V(L), \text{synsets}, \text{Wordnet})$$ (15)

The multisets $\text{NC}(L)$ and $\text{VC}(L)$ contain the noun concepts and verb concepts found in the set of labels $L$, respectively. The multiset $N(F)$ contains the noun facts that map from noun phrases to nouns concepts in YAGO. The multiset $V(F)$ contains the verb facts that map from verbs to verb concepts in Wordnet.

Both relations are one-to-many, which allows multiple concepts to be extracted from a polyseme. Incorrect word senses may be included in the concepts set, which causes the ambiguity. This problem will be solved in a later step.

4.3 Hierarchical Hypernym Graph Construction

A hierarchical hypernym graph HHG is a hierarchical graph of concepts for $L$ that is used during the topic identification process. It consists of two subgraphs. The first subgraph is a hypernym tree HT, which provides a pyramidal view of concepts based on their hypernym relationships and HT is a rooted tree. A vertex is a concept. An edge is a leaf-to-foot HYPERNYM relationship between two concepts. The
Noun HHG

YAGO

Noun HHG

Creation
(1, 0, 1, 1, 3)
(1, 0, -2, 1, 1, 2)

Product
(1, 0, 1, 1, 4)
(1, 0, -1, 2, 1, 2)

Book
(1, 0, 1, 1, 5)
(1, 0, 1, 2, 4, 2)

Notebook
(1, 0, 1, 1, 6)
(1, 0, 6, 10, 10, 5)

Computer
(2, 0, 2, 2, 6)
(1, 5, 6, 6, 10, 5)

Calculator
(1, 0, 1, 1, 6)
(1, 1, 10, 10, 5)

Laptop
(1, 0, 1, 1, 7)
(1, 1, 1, 1, 2, 3, 2)

Mac Mini
(1, 0, 1, 1, 7)
(1, 2, 1, 2, 3, 3, 3)

Numpad
(1, 0, 1, 1, 6)
(1, 10, 10, 5)

Calculator and
notebook is needed. (C)

More devices are
required. (F)

Small sensors
are needed. (D)

Signal cables
are wanted. (E)

??? (G)

Input Extraction

Verb Extraction

N(\text{in}(h)) = \text{Notebook:1, Mac Mini G4:1, Calculator:1, Numpad:1, Keyboard:1}

N(\text{out}(h)) = \text{Device:1, Sensor:1}

Fig. 2 In the upper half, the noun hierarchical hypernym graph HHG\_{in(h),out(h)} where h is the concept "???(G)" is shown. The solid line graph is the hypernym tree HT\_{in(h),out(h)}. The dashed line graph is the words to concepts graph CG\_{in(h),out(h)}. In the lower half, the tree representation of the idea graph is displayed on the left hand side. The diagram on the right hand side shows the result of the input extraction \text{in}(h) and \text{out}(h).
all concepts in the set of concepts $X$ existed in ontology $Q$, until the root concept is reached. The relationships sets $H(NC(L), YAGO)$ and $H(VC(L), Wordnet)$ contain all of the hypernym relationships from the leaf concepts to the root concept in the sets of concepts $NC(L)$ and $VC(L)$, respectively. The noun and verb hypernym trees are constructed from the $H(NC(L), YAGO)$ and $H(VC(L), Wordnet)$ relationships sets, respectively. Wordnet has multiple root concepts, so the number of noun HTs can be greater than one, whereas the number of noun HTs is always one.

4.4 Topic Selection Methods

To determine a plausible topic for $h$ in an idea chart $T$, the following methods evaluate each concept in the HT in $(h:\text{out}(h))$ and select one of them as a topic for $h$.

4.4.1 Branch Ratio (BR)

Originally proposed by Lin [5], the branch ratio (BR) selects the concept with the highest generalization power. The BR in the sample hierarchical hypernym graph. The value of BR in the $h$ concept is computed based on the $w_{in(h)}$ of the $Machine$ concept divided by the total $w_{in(h)}$ for the $Machine$, $Keyboard$, and $Sensor$ concepts, i.e., $\frac{1}{2}\cdot\frac{1}{2} = \frac{1}{4}$.

Note that, the leaf concepts always yield $\infty$. A concept that yields a low BR is a suitable topic choice. A lower BR implies a higher generalization power for its children concepts. A concept that is $> 0.68$ is considered insignificant [5] and they are ignored. The concept that yields the lowest branch ratio is selected.

4.4.2 Concept Counting (CC)

Originally proposed by Lin [5], the concept counting (CC) is a simple and efficient topic selection method. CC is the frequency of concept $x$ in in$(h)$, which is defined as follows.

$$CC(x) = \begin{cases} \frac{\text{NC}(\text{in}(h))}{\text{NC}(\text{in}(h))} & \text{if } x \text{ is a noun.} \\ \frac{\text{VC}(\text{in}(h))}{\text{VC}(\text{in}(h))} & \text{if } x \text{ is a verb.} \\ 0 & \text{otherwise} \end{cases}$$

(18)

where $\text{NC}(x) = \sum_{w \in \text{in}(h)} w_{\text{in}(h)}(w)$.

A higher CC is better. The concept with the highest frequency is selected as the topic. Figure 2 shows the CC values for each concept. The CC of the $Device$ concept is 0 because there is no word in N(in$(h)$) is mapped to this concept.

4.4.3 Ratio Balance (RB)

Originally proposed by Tiun et al. [6], the ratio balance (RB) requires that the accumulated weight is higher than that expected for each level. The RB of concept $x$ is defined as follows:

$$RB(x) = w_{\text{in}(h)}(x) - \frac{w_{\text{in}(h)}(\text{RB})}{d(x)}$$

(19)

where the depth function $d(x)$ returns the depth of $x$ in HT (the depth of a root concept is one). RB is better. The concept with the highest RB is selected as the topic. Figure 2 shows the RB values for each concept. The RB of the $Device$ concept is computed from its $w_{\text{in}(h)}$ value based on the expected value, which is computed from the $w_{\text{in}(h)}$ of the $Entity$ concept divided by the depth of the $Device$ concept, i.e., $5 - \frac{5}{4} = \frac{15}{4}$.  

4.4.4 Harmonic Mean (HM)

All previous methods have identified topics based on a consideration of in$(h)$ alone, which is the set of labels a topic describes. The set of idea labels that are not mentioned in $h$ (i.e., out$(h)$) have not been used before. The out$(h)$ set provides clear boundaries for the topic of $h$. We propose that the harmonic mean (HM) can be used to evaluate each topic based on a consideration of the in$(h)$ and out$(h)$ sets.

Given a concept $x$, the adjusted weight of the concept $x$ is defined as follows.

$$aw_x = w_{\text{in}(h)}(x) - \alpha w_{\text{out}(h)}(x)$$

(20)

The constant $\alpha$ is the penalty rate that indicates the degree of deduction for each occurrence of words in out$(h)$ that map to the children of the concept $x$ or concept $x$ itself.

A concept in a hierarchical hypernym tree with a high adjusted weight and a high depth is considered to be a plausible topic because it covers a high proportion of the words in in$(h)$ compared with the number of words in out$(h)$ and it has high specificity. HM considers the adjusted weight and
Fig. 3 The pruned hierarchical hypernym graph (original is Fig. 2) used to extract the second topic.

the depth of a concept equally, which can be interpreted as the weighted average of the adjusted weight and depth.

Given the concept $x$, HM of $x$ is defined as follows.

$$HM_n(x) = \frac{2 \times \text{naw}_0(x) \times \text{nd}(x)}{\text{naw}_0(x) + \text{nd}(x)}$$

(21)

where

- $\text{naw}_0(x) = \frac{aw_0(x) - \min_{aw_0(m)} aw_0(m)}{\max_{aw_0(m)} - \min_{aw_0(m)}}$
- $\text{nd}(x) = \frac{\text{nd}(x) - \min_{\text{nd}(m)}}{\max_{\text{nd}(m)} - \min_{\text{nd}(m)}}$

HM is undefined for negative numbers, so the weight and depth need to be normalized in the range [0, 1]. The HM values with three different $\alpha$ values ($\alpha = 0$, $\alpha = 0.5$, $\alpha = 1$) are shown in Fig. 2. To compute $HM_{0.5}(\text{Device})$, we compute $\text{naw}_{0.5}(\text{Device}) = \frac{4 - (-0.5)}{4 - (-0.5)} = 1$ and $\text{nd}(\text{Device}) = \frac{4 - 1}{4} = \frac{3}{4}$. Thus, $HM_{0.5}(\text{Device}) = \frac{2 \times 1 \times \frac{3}{4}}{1 + \frac{3}{4}} = \frac{2}{3}$ is yielded. A concept with a high HM is plausible. The concept with the highest HM is selected.

4.5 More Non-overlapping Topics

Any text usually mentions one or more topics. To identify multiple topics, the most naive approach is to select the top-$n$ concepts evaluated by a topic selection method. As shown in Fig. 2, the first and the second ranks are Machine and Device based on the identification of the first and second topics using HM1. Both concepts are similar and Device is the hypernym of Machine, so Device should not be chosen as the second topic. Instead, another concept should be selected that does not overlap with the previously selected topics.

We follow the method proposed by Lin[5]. After the first topic is selected, we extract the next most significant non-overlapping topic, which does not share any commonalities with the first topic, by removing all of the words in

the CG that refer to the children of the first topic in the HT. Next, the HHG is re-constructed and we re-evaluates every concept using the same topic selection method. The best concept in the new graph is chosen as the second topic. This process can be repeated to select the third topic, the fourth topic, and so on. Figure 3 shows the pruned HHG produced by extracting the second non-overlapping topic from the original HHG (Fig. 2). The Keyboard is chosen as the second topic using the HM1 method.

This process performs word sense disambiguation for polysemes simultaneously. Figure 2 shows that the “Notebook” word has two word senses, i.e., Laptop and Notebook. Laptop is much more likely to be the correct sense because the majority of the content relates to the Machine. This process removes all other senses (e.g., Notebook), after the correct sense (e.g., Laptop) is selected.

5. Empirical Evaluations

The proposed topic selection methods (HM0, HM0.5, and HM1) and existing topic selection methods (BR, CC, and RB) were evaluated empirically using idea charts produced by two creativity techniques, i.e., the KJ Method and mind mapping. An automatic topic identification system was implemented for the evaluation. The POS tagging and lemmatization of nouns and verbs during the extraction step were performed using Stanford CoreNLP\(^1\).

5.1 KJ Method

The experiments were conducted with 87 participants who were Masters or Ph.D. students and who understand the KJ Method well. We developed a web application that allowed participants to construct an idea chart (KJ chart) using their computer. Participants were asked to create at least one idea chart for any problem of interest.

The participants input their ideas into the system to create virtual labels and they performed drag-and-drop operation between two labels to form a new group. When a group was formed, a blank group header was automatically created. When a blank group header was clicked, the tree representation of the current chart and the position of the clicked group header were fed into the system. Next, automatic topic identification was performed.

Lists of pairs of the first and the second identified topics produced by the five selection methods (as described in Sect.4.4) were displayed on the screen. The order of pairs was random. Participants were asked to rank these pairs based on their relevance to the group’s content. The list of noun topics was displayed first, followed by the list of verb topics. After ranking, the participants were asked to input a correct noun topic and a correct verb topic to assess their opinions and measure the system satisfaction. The results of the ranking and system satisfaction assessment were analyzed to evaluate the five topic selection methods.

\(^1\)http://nlp.stanford.edu/software/corenlp.shtml
In total, 1,934 labels and 583 group headers are drawn in 112 idea charts by the participants. The average numbers of HTs in an HHG were 1 in noun topics and 24.89 in verb topics. The average depth of HTs was 5.12 and 1.56 for noun and verb topics, respectively. The average number of concepts found in the HTs was 97.80 and 86.17 for the noun and verb topics, respectively. The average numbers of HTs in an HHG were 1 in noun topics and 24.89 in verb topics.

Based on the preliminary requirement that $BR \leq 0.68$, some group headers with no concepts satisfied this requirement. To ensure fair comparisons, these groups were removed from the evaluation. Only 339 and 373 group headers remained for evaluation in the noun and verb topics, respectively.

### 5.1.1 Topic Ranks

Table 1 shows the ranking results produced by the five topic selection methods for the noun and verb topics (lower results were better). Wilcoxon’s rank sum test [15], which is a non-parametric statistical hypothesis test, was used to test whether the rank results for each topic selection method were significantly different. The performance with $BR$ was the baseline. Any ranks were significantly better than the baseline rank are marked with the * symbol. $HM_{0.5}$ and $HM_1$ with significantly higher ranks than $HM_0$ are tagged with the + symbol. Wilcoxon’s rank sum test was used to compare all of the experimental results. * and + symbols are used throughout the results. Note that RB is absent from some experiment results due to a technical issue.

### 5.1.2 Suggestion Accuracy

The correct noun and verb topics input by participants were compared with the identified topic pairs selected by the five topic selection methods. For Top-1, only the first concept of the identified topic pair was compared with the correct topic. For Top-2, the concatenation of the first and the second concepts was compared with the correct topic. The Porter stemmer [16] was used to stem all words before making comparisons. If at least one word was the same, the identified topic pair was suggested correctly. Table 2 shows the suggestion accuracy with the five topic selection methods.

### 5.1.3 Average Word Similarity

Instead of comparing the suggestion accuracy directly, we computed the average word similarity score of the Cartesian product of the set of all words in the first/second identified concepts and the set of all words in the correct topic. Lin’s word similarity algorithm [17] was applied. For Top-1, only the first identified concept was computed. For Top-2, the average score of both the first and the second identified concepts was computed. The average score is ranged from 0 to 1. Table 3 shows the average similarity scores with the five topic selection methods.

### 5.2 Mind Mapping

IEICE TRANS. INF. & SYST., VOL.E96–D, NO.1 JANUARY 2013
files. Xmind.net Share\(^1\) is a mind mapping sharing website that makes approximately 40,000 idea charts (mind maps) publicly available for download. In the experiment, only English idea charts were downloaded. We extracted the tree representation, the positions of all group headers, and the content written in all of the group headers in each idea chart. The tree representation and the positions of all group headers were input into our implemented system and the automatic topic identification process was performed.

The pairs of the first and the second predicted topics identified using the five topic selection methods were output and compared with the content of the group headers. The suggestion accuracy and the average word similarity used in the KJ Method experiment were also applied with the same settings in this experiment.

We extracted 339,765 labels and 197,505 group headers from 9,575 English idea charts. Because of the preliminary requirement that BR \(\leq 0.68\), the group headers were reduced to 96,468 and 62,378 for nouns and verbs, respectively. Table 4 shows the suggestion accuracy using the five selection methods in the mind mapping experiment. Table 5 shows the average word similarity scores for the five selection methods in the mind mapping experiment.

### 6. Discussion

The experimental results showed that, CC, RB, and HM delivered significantly higher topic identification performance than BR (\(^*\) symbols). The HM\(_{0.5}\) and the CC yielded the best rank evaluations by participants during noun and the verb topic identification, respectively. CC yielded the highest suggestion accuracy in most cases, while HM\(_{0.5}\) yielded the highest average word similarity score in most cases. The inclusion of the set of negative labels in the calculations (HM\(_{0.5}\) and HM\(_{1}\)) delivered significant higher performance compared with when it was excluded (HM\(_{0}\)) from verb topics identification (\(^\alpha\) symbols).

The performance of HM was not significantly different from CC, but HM had some advantage over CC. HM could discover topic concepts that were represented by words did not appear in the set of labels. In our opinion, these concepts were more plausible than the concepts in the most frequent words because they are the generalizations of idea groups using different words.

In contrast to BR, which never considered leaf concepts or any concepts that yield BR > 0.68 in the HHG as topic candidates, HM lacked these requirements. The amount of text per idea group is usually short in idea visualization programs. Most are simple phrases. Using a small amount of extracted concepts, BR failed to make any suggestions in most cases, whereas HM always managed to generate a suggestion even when only one concept was found.

### 7. Conclusion

This paper proposed a novel selection method, harmonic mean, for knowledge-based automatic topic identification, which was based on Lin’s framework. HM is designed to identify topics in extremely short texts extracts, which is usual in the idea visualization programs. HM was compared with BR, which was proposed by Lin, and CC, which is a powerful statistical-based method.

Two experiments were performed using datasets from two idea visualization programs, i.e., the KJ Method and mind mapping. CC and HM significantly outperformed BR in terms of performance and consistency. CC and HM differed little in terms of performance, but HM discovered different plausible topics that did not appear in the text, whereas CC could not.

### Acknowledgments

This research was supported by the Japan Advanced Institute of Science and Technology (JAIST) under the Graduate Research Program (GRP).

---

\(^1\)http://www.xmind.net/share/
References

[1] J. Kawakita, “The KJ method: A scientific approach to problem solving,” Tech. Rep., Kawakita Research Institute, Tokyo, 1975.
[2] L.F. Young, “Idea processing support: Definitions and concepts, chapt. 8,” Decision Support and Idea Processing Systems, Wm. C. Brown Publishers, pp.243–268, 1988.
[3] G. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. Miller, “Introduction to wordnet: An on-line lexical database,” Int. J. Lexicography, vol.3, no.4, pp.235–244, 1990.
[4] F. Suchanek, G. Kasneci, and G. Weikum, “Yago: A core of semantic knowledge,” Proc. 16th International Conference on World Wide Web, pp.697–706, 2007.
[5] C.Y. Lin, “Knowledge-based automatic topic identification,” Proc. 33rd annual meeting on Association for Computational Linguistics, ACL ’95, Stroudsburg, PA, USA, pp.308–310, Association for Computational Linguistics, 1995.
[6] S. Tiun, R. Abdullah, and T. Kong, “Automatic topic identification using ontology hierarchy,” Computational Linguistics and Intelligent Text Processing, pp.444–453, 2001.
[7] X. He, C. Ding, H. Zha, and H. Simon, “Automatic topic identification using webpage clustering,” Proc. IEEE International Conference on Data Mining, 2001, ICDM 2001, pp.195–202, IEEE, 2001.
[8] H. Özmutlu, F. Çavdur, S. Özmutlu, and A. Spink, “Neural network applications for automatic new topic identification on excite web search engine data logs,” Proc. American Society for Information Science and Technology, vol.41, no.1, pp.310–316, 2004.
[9] C. Clifton, R. Cooley, and J. Rennie, “Topcat: Data mining for topic identification in a text corpus,” IEEE Trans. Knowl. Data Eng., vol.16, no.8, pp.949–964, 2004.
[10] K. Coursey, R. Mihalcea, and W. Moen, “Using encyclopedic knowledge for automatic topic identification,” Proc. Thirteenth Conference on Computational Natural Language Learning, pp.210–218, Association for Computational Linguistics, 2009.
[11] O. Muñoz-García, A. García-Silva, O. Corcho, M. de la Higuera Hernández, and C. Navarro, “Identifying topics in social media posts using dbpedia,” Proc. NEM Summit 2011, pp.81–86, 2011.
[12] D. Leake, A. Maguitman, and T. Reichherzer, “Topic extraction and extension to support concept mapping,” Proc. Sixteenth FLAIRS, 2003.
[13] K. Viriyayudhakorn, S. Kunifuji, and M. Ogawa, “A comparison of four association engines in divergent thinking support systems on wikipedia,” Knowledge, Information, and Creativity Support Systems, pp.226–237, 2011.
[14] J. Villalon and R. Calvo, “Concept extraction from student essays, towards concept map mining,” Ninth IEEE International Conference on Advanced Learning Technologies, 2009, ICALT 2009, pp.221–225, 2009.
[15] F. Wilcoxon, “Individual comparisons by ranking methods,” Biometrics Bulletin, vol.1, no.6, pp.80–83, 1945.
[16] M. Porter, “Snowball: A language for stemming algorithms,” 2001.
[17] D. Lin, “Automatic retrieval and clustering of similar words,” Proc. 17th international conference on Computational linguistics-Volume 2, pp.768–774, Association for Computational Linguistics, 1998.

Kobkrit Viriyayudhakorn is a doctoral candidate at Japan Advanced Institute of Science and Technology (JAIST). His research interest is computational creativity such as creativity support systems and creative autonomy.

Susumu Kunifuji was born in 1947. He received B.E., M.E., and D.E. degrees from the Tokyo Institute of Technology in 1971, 1974, and 1994, respectively. He worked as a researcher at the International Institute for Advanced Study of Social Information Science, FUJITSU Ltd. (1974-1982), Chief researcher at the Institute for New Generation Computer Technology (1982-1986), Manager of International Institute for Advanced Study of Social Information Science, FUJITSU Ltd. (1986-1992), Professor of School of Information Science at JAIST (1992-1998), and Director of Center for Information Science at JAIST (1992-1998). He is currently a professor at the School of Knowledge Science, Japan Advanced Institute of Science and Technology. He is a member of JSAI, IPSJ, SICE, and JCS among others.