Future Generation Computer Systems 76 (2017) 510–518

Contents lists available at ScienceDirect

Future Generation Computer Systems

journal homepage: www.elsevier.com/locate/fgcs

An optimized approach for massive web page classification using entity similarity based on semantic network

Huakang Li [a], Zheng Xu [b,*], Tao Li [a], Guozi Sun [a], Kim-Kwang Raymond Choo [c]

[a] Key Lab of Big Data Security and Intelligent Processing, Institute of Computer Technology, School of Computer Science & Technology, School of Software Nanjing University of Posts and Telecommunications, Nanjing, 210023, China

[b] The Third Research Institute of the Ministry of Public Security, Shanghai, 201204, China

[c] Department of Information Systems and Cyber Security, The University of Texas at San Antonio, San Antonio, TX 78249-0631, USA

HIGHLIGHTS

- A weight estimation algorithm based on the depth and breadth of Wikipedia network is used to calculate the class weight of all Wikipedia words.
- A kinship-relation association based on content similarity was proposed to optimize the unbalance problem.
- Bayesian classifier is used to estimate the page class probability.

ARTICLE INFO

Article history:
Received 16 February 2016
Received in revised form
11 December 2016
Accepted 1 March 2017
Available online 21 March 2017

Keywords:
Web page classification
Semantic network
Kinship-relation association
Entity class probability
Hereditary weight

ABSTRACT

With the development of mobile technology, the users browsing habits are gradually shifted from only information retrieval to active recommendation. The classification mapping algorithm between users interests and web contents has been become more and more difficult with the volume and variety of web pages. Some big news portal sites and social media companies hire more editors to label these new concepts and words, and use the computing servers with larger memory to deal with the massive document classification, based on traditional supervised or semi-supervised machine learning methods. This paper provides an optimized classification algorithm for massive web page classification using semantic networks, such as Wikipedia, WordNet. In this paper, we used Wikipedia data set and initialized a few category entity words as class words. A weight estimation algorithm based on the depth and breadth of Wikipedia network is used to calculate the class weight of all Wikipedia Entity Words. A kinship-relation association based on content similarity of entity was therefore suggested optimizing the unbalance problem when a category node inherited the probability from multiple fathers. The keywords in the web page are extracted from the title and the main text using N-gram with Wikipedia Entity Words, and Bayesian classifier is used to estimate the page class probability. Experimental results showed that the proposed method obtained good scalability, robustness and reliability for massive web pages.

© 2017 Elsevier B.V. All rights reserved.

1. Introduction

With the development of mobile technology, the users browsing habits are gradually shifted from only information retrieval to active recommendation. The classification mapping algorithm between user interests and web contents has been become ever more complicated in the volume and variety of web pages. Web page classification plays a vital status on the Internet information management, convenient retrieval, web page crawling and user profile identification [1,2]. In order to increase the friendly browsing and rapid retrieval experience, the directories of some sites, such as Yahoo![1] and DMOZ ODP,[2] define some series site structure based on web page content information to strengthen the structure and hierarchical browsing approaches. ODP employed over 78940 people to engage the web page classification for page maintenance in

* Corresponding author.
E-mail address: xuzheng@shu.edu.cn (Z. Xu).

http://dx.doi.org/10.1016/j.future.2017.03.003
0167-739X/© 2017 Elsevier B.V. All rights reserved.

[1] http://www.yahoo.com/.
[2] http://www.dmoz.org.

Netscape Communications Corporation 2008 report.[3] With the development of mobile, the company increased ten million annually and the average web page numbers of these Web sites almost reach to millions lever. The web page classification depending on manpower editing becomes increasingly quality verification and other tedious processes of web pages do not meet demand for the rapid expansion of web sites. The web page classification labeling has become consume enormous, human and financial costing for Internet applications. Some automated or semi-automated mechanisms for Web page classification with high accuracy, reliability and scalability have an obligation to replace the man powered editing.

Traditionally, web page classification systems [2] are almost based on supervised machine learning algorithm, which estimates a labeled classification training model to forecast the testing data. Web page classification methods can be divided into several main categories: subject classification, functional classification, sentiment classification and other forms of classification.

Subject classification is the theme for page content classification [3], which is more conducive to site information management and publishing. Current integrated information sites almost use this classification system like Yahoo!, American Online, and Sina, have the channels named "health", "culture", "travel" and etc. The web pages are, firstly, encoded with multi-dimensional vector representation for facilitating machine learning and classification processes. The dimensionality reducing methods or feature selection algorithms are used to save both time and space for computation. Then, machine learning methods, like SVM [4], ANN [5], Rocchio [6] and Bayesian Classifier [7], are applied for classifying web pages.

Actually, the data set used by the conventional methods has significant structural characteristics. The accessible data set provider will optimize the data structure to correspond the customer's requirement. On the other hand, the users have sufficient resources and capabilities to integrate the data structure and optimization for a small data collection. However, it is impossible to do these things when the data are too cumbersome for the users. At the same time, the data set may be brought from different providers whose data structures are skimble-scamble. Further more, the traditional data analysis and processing methods are too complex and inefficient to accommodate the diverse circumstances noises when the page data reach TB/PB level. Big data have brought a lot of development opportunities and challenges in the content classification field. More and more big news portal sites and social media companies use the computing servers with larger memory to deal with the massive document classification, based on traditional supervised or semi-supervised machine learning methods.

This paper proposed web page content classification algorithm based on Wikipedia knowledge with network topology. Wikipedia encyclopedia knowledge is the largest knowledge base all over the world, and has over 200 different language knowledge bases with thousands or even millions of entries. These knowledge bases are entries every day from the separate group to expand, editing and finishing. We defined the Preliminary Classes (PCs) space on web page content consulting various category systems of portal sites at first. Then, some category words in Wikipedia Knowledge Network (WKN) were defined as Elemental Keywords (EKs) for PCs. During the Wikipedia category tree breeding, we proposed three inheritance association algorithms: Preliminary Association (PA), Rule Association (RA) and Kinship-relation Association (KRA). Class Probabilities (CPs) of all Wikipedia category words were estimated using Levenshtein distance between the category words and EKs for the association method. RA method analyzed the breadth and depth between Wikipedia category word and father or ancestor words to estimate the CPs of the current category words. KRA method introduced the sister concepts to optimize the hereditary weights of Wikipedia category words for multiple father's conditions.

In specific page classification processing, we collected several typical Chinese web pages as test samples. And a Main-text founder process (MFP) tool was used to extract the kernel content information. The contents were segmented into words with minimum phrases using a Chinese tokenizer. And these words were combined to obtain additional vocabulary bag using N-gram algorithm referring to Wikipedia categories phrases. We used a naive Bayes classifier model to estimate the page PCs according to the word PCs. A sigmoid function was put in place to optimize the hereditary weights considering to the unbalance of word frequency and word length.

The experimental results showed that KRA method based on WKN obviously enhanced the classification accuracies for all benchmarks. The proposed method had high accuracy, reliability and scalability for a variety of page qualities comparing with traditional BOW and TF–IDF methods. According to the update model of WKN, we alleged that the KRA algorithm based on WKN was suitable for web page classification, especially for large data mining.

The paper will be presented as following: Section 2 introduced the recent web page classification methods. The motivation of using Wikipedia knowledge database and the reconstruction algorithm were explained in Section 3. Section 4 presented the kernel of web page keyword extraction and classification approach. Finally, the experimental evaluation was discussed in Section 5. Conclusion and feature works were given in Section 6.

2. Related work

Web pages are classified based on their contents or subjects in the topic-based classification research. For the English document, it is very easy to segment a document into a vector of words. And each word indicates a certain concept in the document of the basic hypothesis. Finally, a web page is represented by a vector of words with weights. This if often referred to as Bag-of-Words approach [4,8]. In order to improve the weight generation, Mladenić [9] introduced N-gram for feature word selection. They found that the terms up to 3-gram were sufficient in most classification systems.

One of the most successful term weighting methods is Term Frequency Inverse Document Frequency (TF–IDF) [10], which is obtained by the product of the local term importance and the global term importance. When the term frequency is very high in a document or category, it is a very nice Unique Feature Vector (UFV) [1]. On the other hand, if a term appears in many documents or categories, and the frequency difference is small, it is a good perform of UFV [6]. Another algorithm named "Balanced Term-Weighting Scheme" is proposed for similarity computation between a document which is often required in classification tasks [11].

After features of training web pages have been selected, various machine learning methods and classification algorithm can be applied to induce the classification function. For profile based classifier, a profile for each category is extracted a set of training web pages that has been predefined as examples of the category. After training all categories, the classifiers, such as Rocchio Classifier [12], Support Vector Machine [13], Neural Network Classifier [14], Linear Least Square Fit Classifier [15] etc. are used to classify new web pages. For parameter based classifiers, such as Naive Bayesian Classifier [7], training examples are used to estimate parameters of a probability distribution.

[3] http://news.netcraft.com/.

For the classification quality, Zhang [16] provided a nonparametric approach to improve the classification performance effectively by incorporating correlated information into the classification process. Wang [17] proposed a constrained clustering scheme that makes decisions with consideration of some background information to improve the accuracy of information clustering. Zhang [18] combined supervised and unsupervised machine learning techniques to the meet robustness of classification performance.

One of the most expressive and human readable representation of the learned hypotheses is sets of relation rules because rule learning based classifier is context sensitive classification. Disjunctive Normal Form rule [19], which performs quite well for document classification, is proposed to interpret a disjunction of simple features. Decision Tree [3] is also considered as rule learning based classifiers because it can be converted into a disjunction of conjunction rules.

With an increasing number of web pages, the limitations of these methods are more and more difficult to solve it. Currently, more and more researchers tried to improve the web page classification with large scale data [20]. Wang [21] observed web pages for a same topic from different language usually share some common semantic patterns, though in different representation forms with a novel joint Nonnegative Matrix Trifactorization (NMTF) based on Dual Knowledge Transfer (DKT) approach. Marath [22] partitioned the subcategories of the Yahoo! web directory into five mutually exclusive groups based on the prior probability distribution and machine learning issues associated with such an imbalanced distribution.

At the same time, the association rule classification based on data set has performance comparable to most well know document classifiers. Association Rules Mining [23] is a data mining task aiming at finding out the relationship between items in any data set. More and more association based Web page classification research has been proposed recently in the development of knowledge bases, such as Wordnet [24], Opencyc [25], Thoughttreasure [26], while these words libraries have their own defects. Numerous researches have been proposed with Wikipedia knowledge [27–29] while the results are not so reliable as the traditional methods. We proposed some schemes to improve the classification algorithm based on Wikipedia knowledge base [30–33], and introduce the improved works in this paper.

3. Wikipedia knowledge database model

3.1. Wikipedia structure

Wikipedia is a multilingual, web-based, free-content encyclopedia project operated by Wikipedia Foundation and based on an openly evitable model. Wikipedia is a live collaboration differing from paper-based reference sources in important ways. Wikipedia encyclopedia knowledge is the largest knowledge base all over the world, and has over 200 different language knowledge bases with thousands or even millions of entries.

Its English version contains more than 4.3 million entries currently. It touches almost every imaginable topic, from mathematics to sociology, from politicians to popular products. Some subjects are considered only briefly, others with essays running through several pages. In addition to getting an extensive coverage, Wikipedia incorporates up-to-date knowledge and its articles are written in contemporary English, which makes it an ideal vehicle to help topic identification. Still, it holds the significant drawback that it has a much less formal and sophisticated structure than proper ontology. And unfortunately sometimes even this simplistic structure was not consistent (among others [34,35] make suggestions regarding how to improve the situation). Wikipedia is just an extremely large conceptual repository, both in terms of width and depth (Fig. 1). One category node may have multiple father nodes, and link with many child nodes and concepts.

3.2. Wikipedia category construction

In order to explain the system construction more perspicuous, we gave several definitions here to distinguish some notion confusing.

- Wikipedia Entity Words (WEW): The vocabulary of every node in WKN. If a node had child nodes, we defined the vocabulary as category entity word which was male gender with genetic properties. If a node was a leaf node or without any additional child nodes, we defined it as a concept word which was female gender without generation capability. When a new child node was added to the concept node, the concept became a category entity node.
- Preliminary Classes (PCs): Several basic classes were defined artificially referring to some Web sites, such as "culture", "sports", "finance", etc.
- Elemental Keywords (EKs): Several Wikipedia category words were predefined as the initial keywords for PCs. For examples, "calligraphy" was in "culture" class; "basketball" was one keyword of "sports"; and "finance" class had the word "stock".

We defined the PCs vector space as $\overrightarrow{C}$, which contained N classes: $\overrightarrow{C} = \{c_1, c_2, \ldots, c_N\}$. Assuming each class $c_n (1 \le n \le N)$ had an elemental keyword dictionary $dict_n$, which included M_n WEWs. Therefore, the $\overrightarrow{dict_n} = \{key^n_1, key^n_2, \ldots, key^n_{M_n}\}$ and we obtained the whole dictionary $\overrightarrow{Dict} = \{dict_1, dict_2, \ldots, dict_N\}$ at the initial status;

According to this preliminary keyword dictionary, each WEW would relate to certain classes. We supposed that there were K unmarked words $\overrightarrow{uMW} = \{w_1, w_2, \ldots, w_K\}$ between each other in Wikipedia knowledge base. The association method of kth word was defined as following:

$$MAP_{w_k \leftarrow C} : w_k \leftarrow c_1 : p_{k1}, w_k \leftarrow c_2 : p_{k2}, \ldots, w_k \leftarrow c_n : p_{kn} \quad (1)$$

Where $p_{kn} (1 \le n \le N, 1 \le k \le K)$ presented the PCs of WEW w_k is in class c_n. This probability distribution illustrated the association relation between current WEW and each PC. In fact, association processing was to deal with the above mapping project. And Wikipedia category nodes which completed the projecting were recognized as marked nodes. We proposed three inheritances associated algorithms: PA, RA and KRA in the WKN tree breeding process.

3.3. Association models

(I) *Preliminary association*

Assuming unmarked WEWs $\overrightarrow{uMW}$ and PC dictionary $\overrightarrow{Dict}$ were orthogonal, which meant any word $w_k \in \overrightarrow{uMN}$ would find one same or similar keyword $key^n_k \in \overrightarrow{Dict}$. We used an edit distance function $Sim(w_k, key^n_k)$ to establishing the overlap proportion between WEWs and EKs eliminating stop-words. A threshold value δ was set to judge whether the WEW w_k could directly find certain keywords key^n_k. When the similarity values $Sim(w_k, key^n_k)$ was bigger than δ, the classification frequency CF_n of current WEW w_k in PCs accumulated 1, otherwise 0.

$$CF_n += \begin{cases} 0 & (Sim(w_k, key^n_k) < \delta) \\ 1 & \text{otherwise.} \end{cases} \quad (2)$$

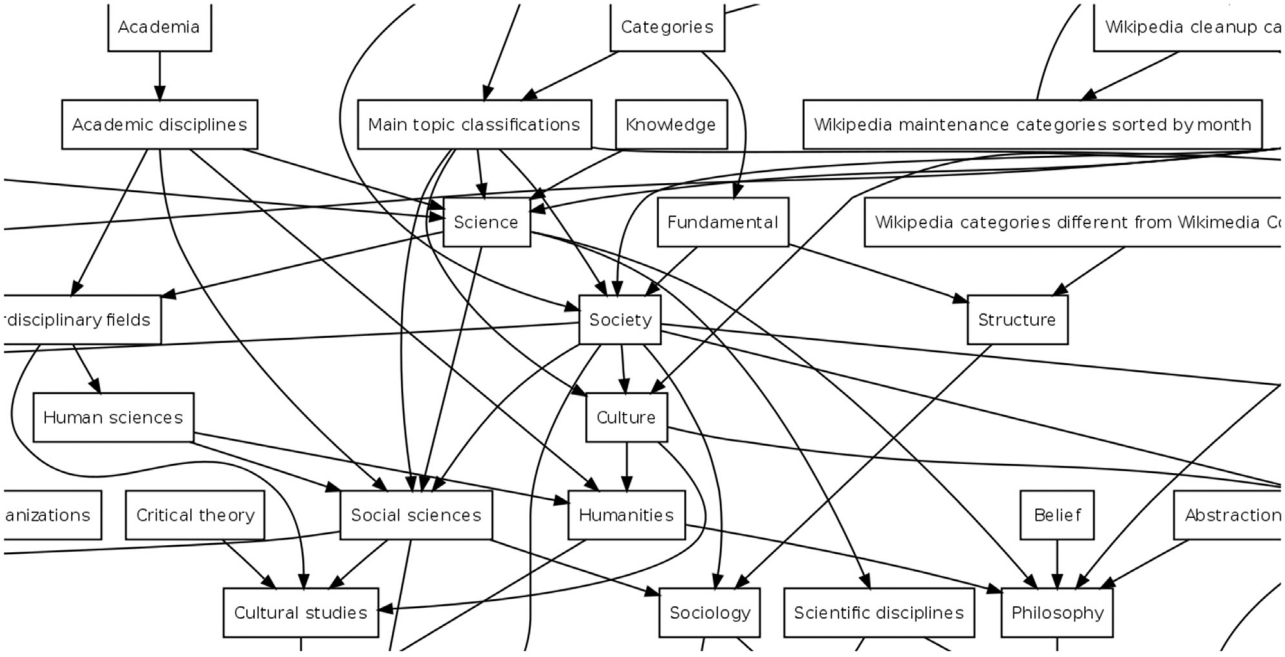


Fig. 1. Simple structure of Wikipedia category network.

Then, a normalization method was used to draw the probability p_{kn} for the WEW w_k in PCs C_n.

$$p_{kn} = \frac{CF_n}{\sum_{n=1}^{N} CF_n}. \tag{3}$$

For example, "Shanxi Province (w_1)" could not find any similar keyword in $\overrightarrow{Dict}$ that its mapping association was $\{w_1 \leftarrow other : 0\}$. The Wikipedia category word "turbo (w_{10})" only matched one word in PCs "mechanical (c_1)". Therefore, the probability of "turbo" represented as $\{w_{10} \leftarrow c_1 : 1, others : 0\}$. Wikipedia category word "chip frequency (w_{40})" matched two keywords in class "computer (c_2)" and "mobile phone (c_3)", which meant the probability distribution of "chip frequency"' was $\{w_{40} \leftarrow c_2 : 0.5, w_{40} \leftarrow c_3 : 0.5, \text{ others} : 0\}$.

For the normal circumstance, the EKs in dictionary $\overrightarrow{Dict}$ were far less than Wikipedia category words. Therefore, a lot of words were not in a position to complete the marking processing. At the same time, expressive variety in China for one thing made it difficult for marking. For instance, the Wikipedia category word "Xinnian (new year)" could not find the keyword "Chunjie (new year)" in PCs "culture" using text similarity directly.

(II) *Rule association*

RA formulated a series of inheriting, and spread rules to evolve the PCs of each WEW in WKN topology with the EKs. We took a top-down inheritance pattern, which meant the probability of one Wikipedia category node relied on the already marked ancestor nodes. If direct fathers were in $\overrightarrow{uMW}$, the algorithm would progressively trace to be recursive until the marked ancestor nodes or root nodes. We made all root nodes were marked and there were numerous Wikipedia category nodes marked in the entire network topology. Thus, all of unmarked nodes could find non-empty ancestor nodes. There were three kinds of generation situations for the RA processing: Single father; multi-fathers in single layer; multi-level and multi-fathers (Fig. 2).

- Single-father: Node X has one marked ancestor node A with PCs $\overrightarrow{P_A}$. Since X cannot get inherited from other ancestral nodes, the PC of X generates directly from A. The definition of association rule:

$$MAP_{w_X \leftarrow w_A} : \overrightarrow{P_X} = f(A \rightarrow X) \cdot \overrightarrow{P_A}. \tag{4}$$

- Multi-fathers in Single-layer: Wikipedia category nodes have multiple father nodes $\overrightarrow{A}$ and these father nodes are EKs or temporary calculated nodes $\overrightarrow{A} = \{A_1, A_2, \ldots, A_T\} (2 \le T \le K)$. Thus, the association rules were defined as:

$$MAP_{w_X \leftarrow w_A} : \overrightarrow{P_X} = \sum_{t=1}^{T} \frac{f(A_t \rightarrow X)}{T} \overrightarrow{P_{A_t}}. \tag{5}$$

- Multi-level and Multi-fathers: Unmarked node X had T direct fathers $\overrightarrow{F} = \{F_1, F_2, \ldots, F_T\}$ while they were not EKs or template marked nodes. The links from fathers F_t to the marked nodes A_t were all single father rule with depth $d_t (1 \le t \le T)$. We considered the mutation rate was increasing as the number of iterations. Therefore, for the unmarked node in multi-level and multi-fathers condition, the PCs were estimated as follows:

$$MAP_{w_X \leftarrow w_A} : \overrightarrow{P_X} = \sum_{t=1}^{T} \frac{f(A_t \rightarrow X) \frac{1}{d_t}}{\sum_{t=1}^{T} 1/d_t} \cdot \overrightarrow{P_{A_t}}. \tag{6}$$

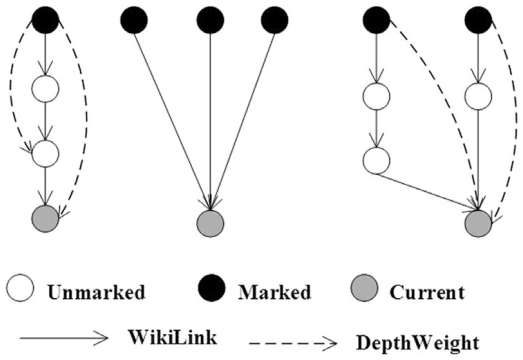


Fig. 2. Rule association processing for WKN.

Table 1
Kinship-relation sample like "Game Type".

Category			Game Type		
Parents	Sisters				
Game	Paronomasia	Mental game	Electronic game	Game book	etc.
Type	Type	Panic			

Table 2
Inherit weights using father and kinship-relation associations.

Fathers	Raw	S_f	S_{f+s} (Leven)	S_{f+s} (hash)
Type	1	0.5	0.25	0.27
Game	1	0.5	0.75	0.73

(III) *Kinship-relation association*

The relationships between category and concept nodes in Wikipedia knowledge base are so complex that some category nodes have many fathers category nodes. Therefore, some research introduced the similarity to improve the relationship with different fathers [36,37]. In the Chinese words similarity calculation, Levenshtein distance and Simhash algorithms are more common to be used. Simhash [38] is practically useful to identify near-duplicates in web documents belonging to a multi-billion page repository. Levenshtein distance is a string metric for measuring the difference between two sequences. And the phrase edit distance is often used to refer specifically to Levenshtein distance [39].

In our method, we used the entity similarity between the current node and direct father node to estimate the hereditary weight caused by the serious imbalance of generations. If the similarity function between estimating node and its father was $Sim(X,\ F)$, the inheritance weighted mapping model was written as:

$$MAP_{w_X \leftarrow w_A} : \overrightarrow{P_X} = \sum_{t=1}^{T} \frac{f(A_t \rightarrow X) Sim(X, F_t)/d_t}{\sum_{t=1}^{T} Sim(X, F_t)/d_t} \cdot \overrightarrow{P_{A_t}}. \tag{7}$$

For some combined-words such as "game type" whose direct fathers were "type" and "game", the generated hereditary weight would be {type : 0.5; game : 0.5} in Table 1. However, in our hominid, "game type" was much closer to "game" than type". We found that each category entity node contained several concepts, which had good text similarity with subcategory nodes. Therefore, we introduced the kinship-relation to separate the difference between father nodes. The new hereditary weight function was estimated as follows:

$$\widehat{Sim} = \alpha * Sim_{father} + \beta * \frac{1}{S} \sum_{s=1}^{S} Sim^s_{sister}. \tag{8}$$

Where Sim_{father} was the word similarity between category entity node and one father. The Sim^s_{sister} shown the word similarity with *sth* sister. S was the concepts' number of the direct father. α and β was constant weight set with 1 and 2 since we thought the sisters' similarity was more import to judge whether this category node betrays this father. The similarity results of "Game type" with fathers "game" and "type" are listed in Table 2.

4. Web page classification

4.1. Keyword extraction

The traditional pure text message contains only a title and body content. Generally web page contains site structure information, such as tags, header, legal statements, billboards and image sources. We crawled some pages from typical Chinese Web sites. Then, a main-text founder processing tool was set up to extract the kernel content of web page. The contents were separated into elementary strings with minimum phrases based on "Jieba" toolbox.

Table 3
Words segmentation case based Wikipedia concepts.

Sentence	Game contains DOS game.	
Split	Words	Term Frequency
Tradition	game	2
	contains	1
	DOS	1
Wiki-based	game	1
	contains	1
	DOS game	1

The thesaurus derived from Wikipedia contains a list of concepts. For each web content in a given corpus, we retrieved the Wikipedia concepts mentioned in the page content. Such concepts were called candidate concepts for the corresponding web page. When trying to find candidate concepts, we adopted an exact matching strategy, by which only the concepts that explicitly appeared in a web page became the candidate concepts. (If a m-gram concept was contained in a N-gram concept (with $n > m$), only the last one became a candidate concept.)

We then constructed a vector representation of a web page, which contained two parts: terms and candidate concepts. For example, considered text fragment "games contain DOS game". Table 3 shown the traditional BOW term vector for this text fragment (before and after stemming), where feature values corresponded to term frequencies. We found that, for each web page, if a word only appeared in candidate concepts, it would not be chosen as a term feature any longer.

The Chinese language word is different English. The longer the word is, more preciousness it is, even using N-gram algorithm based on WEWs. Therefore, we assigned each extracted word in web page with a score for the information of words length such as:

$$Score = TF * \log(1 + len(L_W)) \tag{9}$$

where TF denoted how many times the keyword had been in the web page and L_W was the byte length of words.

4.2. Preliminary class estimation

Assuming one web page had Y extracted keywords $Key = \{key_1, key_2, \ldots, key_Y\}$, and the probability of keyword key_y in nth PCs was p_{ny}. We used a Naive Bayes model to estimate the class probability (CP) $p(c_n|P)$ of the current web page.

$$p(c_n|P) = \sum p(c_n|key_y) p(key_y|P). \tag{10}$$

Where $p(c_n|key_y)$ is the CP of key_y, and $p(key_y|P)$ is able to be obtained by Bayes function:

$$p(key_y|P) = \frac{p(P|key_y)p(key_y)}{p(P)} = \#P \cdot \frac{tf_y}{TF}. \tag{11}$$

Here tf_y was the term frequency of key_y in web page P. TF showed that the total team frequency of all *keys* $\#P$ was the number of all web pages. Because the TF and $\#P$ were same to different web pages, the nth CP was:

$$P_n = \sum_{y=1}^{Y} Score_n \cdot p_{ny}. \tag{12}$$

When the web page CPs were obtained, we taken the largest probability class as the final PCs of the current web page:

$$P \leftarrow c_z, z = argmax_n(P_n, n). \tag{13}$$

Table 4
Elemental keywords number for each class.

Class	Word	Class	Word
Cult	151	Health	148
Edu	95	Religion	66
Ente	91	Science	195
Finc	170	Sport	46
Gov	137	Travel	57

Actually, Page PCs distinguish was not significant if keyword PCs were assigned to two different class equally [40]. We assumed that the page CPs should be described as an inverse sigmoid function of the keyword CPs. When the keyword CP closed to 0 or 1, page PCs grown or dropped fast. Page PCs increased slowly when keyword PCs were near 0.5.

$$P_n = \sum_{y=1}^{Y} Score_n \cdot \widehat{p_{ny}}. \tag{14}$$

Where,

$$p_{ny} = \frac{1}{1 + e^{-\lambda(\widehat{p_{ny}}+\varphi)}}. \tag{15}$$

The sigmoid function tended to be linear when λ was very small. When λ was large, the start and end areas were sharp and the middle area was very stable. φ is the slope of Sigmoid function for fine tuning.

5. Experiment and evaluation

5.1. System structure

The process of the whole system was very simple. First, we reprocessed the downloaded Wikipedia knowledge base and saved into a database. Then, EKs dictionary was initialized based on "sogo" topic dictionary and Wikipedia knowledge base. The CPs of all WEWs were generated using PA, RA and KRA. The main contents of crawled web pages were provided by an "MFP" toolbox. Page keywords were selected using a tokenizer refer to WEWs by the *N*-gram applications [41]. After introducing an inverse sigmoid function to balance the keywords probability, we used a simple Bayesian classifier to estimate the final class of the web page. The detail data structure, selection and experimental results were presented below.

5.2. Dictionary initialization

Chinese Wikipedia knowledge base has 23 root categories, and category node is linked with a number of subcategories. Some root category words can represent a PC specifically, such as "mathematics" and "religion" while the CPs of other root nodes are very ambiguous. The root node "technology" not only contains industry, computer and etc., but also includes "horticultural", "media" and "medicine" subcategories. Therefore, we manually checked the top two or three layers of WKN and marked each root node as an elemental keyword in the closest class. The PC probability would be redefined when a subcategory was obviously betrayed his father. Table 4 lists the number of EKs in each class.

5.3. Wikipedia concept filtering

We did not include all category/concept nodes of Wikipedia in the dictionary. Some nodes, such as "List of ISO standards" or "1960s", do not contribute to the achievement of improved discrimination among web pages. Thus, before building the dictionary from Wikipedia, we removed several nodes deemed not useful. To this end, we implemented a few heuristics. All nodes of Wikipedia which belonged to categories associated with chronology, such as "Years", "Decades", and "Centuries", were removed. Table 5 provides a breakdown of the resulting number of elements (terms, concepts, and links) used to build the dictionary.

Table 5
Number of terms, concepts, and relations after filtering.

Terms in Wikipedia corpus	1 207 380
Concept After Filtering	1 190 131
Categories	17 249
Relations in Wikipedia Corpus	5 200 281
Category to Subcategory	264 630
Category to Concept	4 935 651

Table 6
Data set for Wikipedia Classification experiment.

Category	Sina	Ifeng	Yahoo-ch
Culture	1 889	1 939	1 989
Education	1 917	1 856	1 995
Entertainment	1 886	1 948	1 994
Finance	1 960	1 994	1 865
Health	1 996	1 759	1 978
Government	1 969	1 929	1 997
Religion	213	1 863	1 991
Science	1 794	1 889	1 643
Sport	1 854	1 080	1 996
Travel	1 953	1 842	1 996
Total	17 431	18 099	19 444

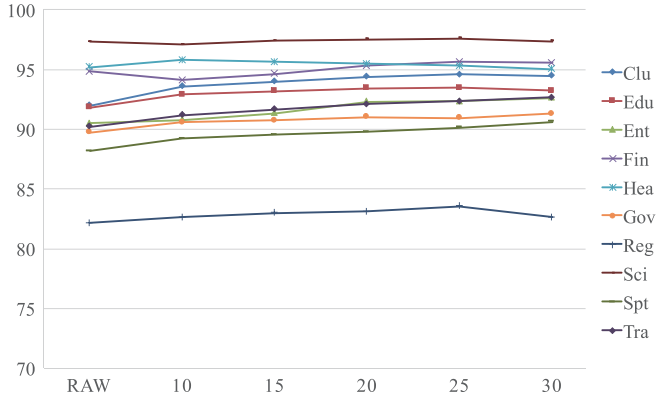
Table 7
Topic prediction results for benchmarks (%).

Cate	Sina		Ifeng		Yahoo-ch	
	S_f	S_{f+s}	S_f	S_{f+s}	S_f	S_{f+s}
Cul	89.57	92.01	92.73	94.79	93.87	97.44
Edu	89.05	91.81	85.61	87.77	79.40	82.66
Ent	88.71	90.51	86.09	87.63	89.37	91.47
Fin	90.77	94.85	90.77	94.73	91.21	95.12
Hea	91.93	95.19	86.47	90.16	85.29	88.78
Gov	85.83	89.74	83.00	83.31	83.68	84.68
Reg	81.22	82.16	82.34	85.13	85.28	86.99
Sci	94.76	97.32	94.81	97.72	96.71	98.11
Spt	86.35	88.19	83.06	84.72	84.12	85.32
Tra	88.12	90.22	88.11	90.88	86.77	89.88
Avg	89.33	92.07	87.52	89.93	87.38	89.86

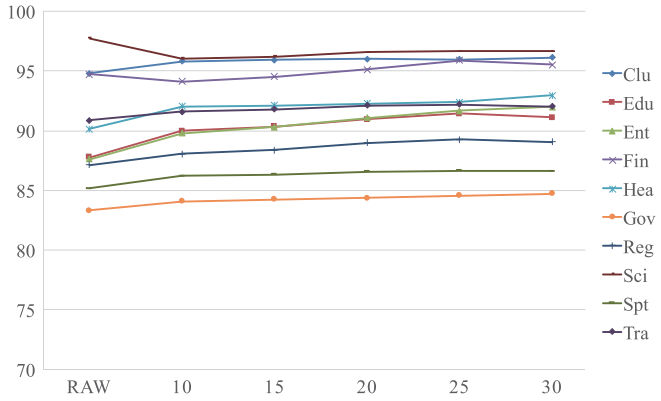
5.4. Data sets for benchmarks

In order to measure the accuracy of the given class prediction model, we selected three representative sites in the top 10 of Chinese portal sites. Table 6 shows the page number crawled with different topics in these sites.

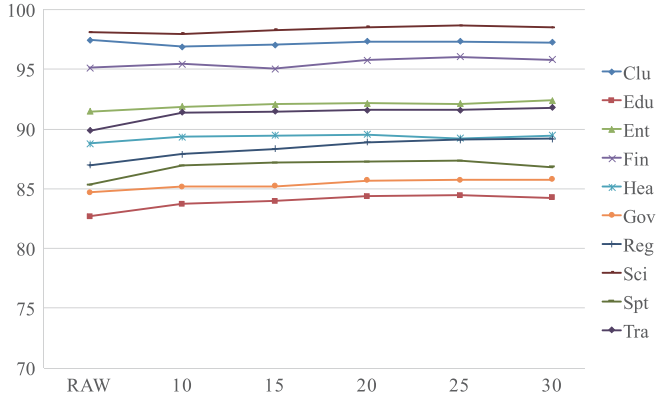
- www.ifeng.com: The PV of normal web browsing is greater than 400 million, and mobile browsing is more than 200 million every day. The user value (UV) is greater than 260 million monthly. The browsing page *PV*/*UV* ratio, average daily effective browsing time indicator is also in the advanced level of Chinese Internet sites. The report from iResearch by 2012 presents that the average income, family economic status, education level, managerial and professional proportion ratio of Ifeng's user are at the top level in China.
- www.sina.com: Sina, which is providing value-added services of Chinese community information is a leading online Media Company. Its registered users are more than 260 million and daily PV is greater than 0 million that arrested the most respected brand in the global community.



(a) Classification of Sina host.



(b) Classification of Ifeng host.



(c) Classification of Yahoo host.

Fig. 3. Classification results for three benchmark with variety φ.

- www.yahoo.com.cn: Yahoo is the world's first portal information website, with operations in 24 countries and regions, providing wide-range of network services for the world's more than 500 million unique users. Yahoo China has over 20 top channels, and Yahoo China web site focuses on information, e-mail, search of the Internet portal based application service for the majority of Internet users.

5.5. Experiment results

Here we would omit to explain the experimental results based on the RA method since several researchers have given some meticulous work before. Table 7 shows that the classification results using the KRA algorithm were increased 2%–3% than simple

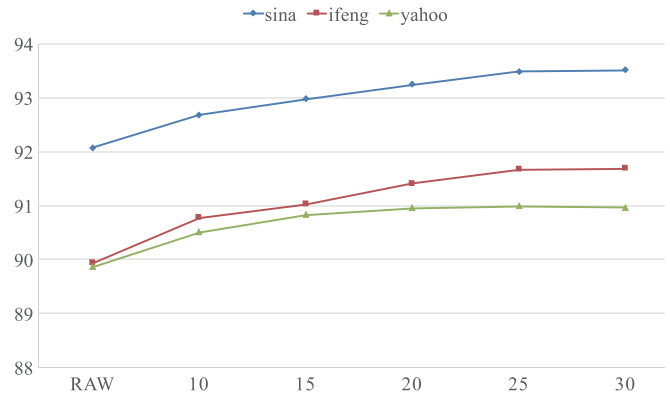


Fig. 4. The total classification result of three benchmark.

similarity models. The most significant performance was the final class whose increasing ratio was almost 4% for all benchmarks. Financial nodes in Wikipedia network had little correlation with nodes in other classes that made the CPs have high cohesiveness. On the other hand, the accuracy of politics and religion were not so significant because the EKs were too less for these two classes. Another reason was the polysemy and ambiguity of politics, and religion nodes were very serious. For instance, "Shaolin Temple", which was one of the most famous temples in China, was virtually introduced as a tourist attraction or cultural relics.

Fig. 3 plots the system accurate variation of all classes with the increasing of the parameter φ for all benchmarks. Generally, the recognition rate is gradually rising φ. When the φ rises to 25, the system accuracy almost reached a peak and when φ is 30 the performances of ifeng and Yahoo! decrease inversely. Therefore, we concluded that the sigmoid function of $\varphi = 25$ settings was the best possible performance for Wikipedia knowledge probability distribution model (Fig. 4).

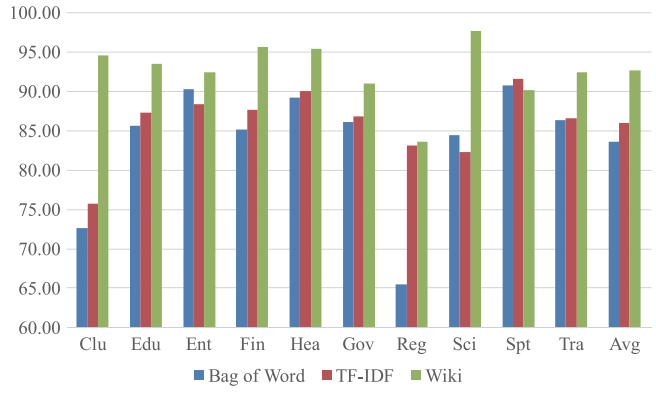
5.6. Discussion

We carried out several contrast experiments to evaluate the performance of classification based on WKN. Fig. 5 shows the performance of BOG, IF–IDF and WKN algorithms on three benchmarks. Obviously, the accuracy of each category in all benchmarks based on WKN is more than 85% while it is not stable using BOG or TF–IDF algorithm. Especially for the culture and technology, the accuracy is almost 95% with WKN since the Chinese Wikipedia knowledge base contains a lot of category nodes associated with these two classes. On the contrary, the accuracy of sports class is no better than BOG or TF–IDF algorithm for all benchmarks. Kernel keywords for sports reports are superstars' name while only limited sport stars nodes are contained in Wikipedia knowledge. For the other categories, the performance of WKN is little better overall compared with BOG or TF–IDF.

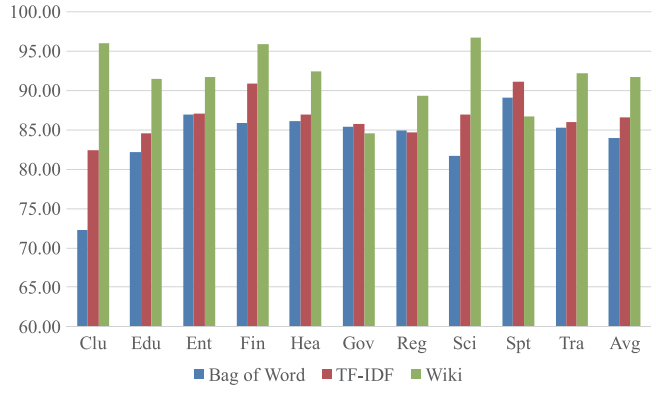
In terms of overall benchmarks, the experimental results based on WKN are obviously better than BOG and TF–IDF in Sina and Ifeng benchmarks. We found that text capacity size from Sina and Ifeng is considerable differences while almost each web page has extensive content information for Yahoo. The accuracy performance based on WKN in every benchmark reaches more than 90%. It is demonstrated that the classification of WKN has good robustness and scalability for different web pages.

6. Conclusions

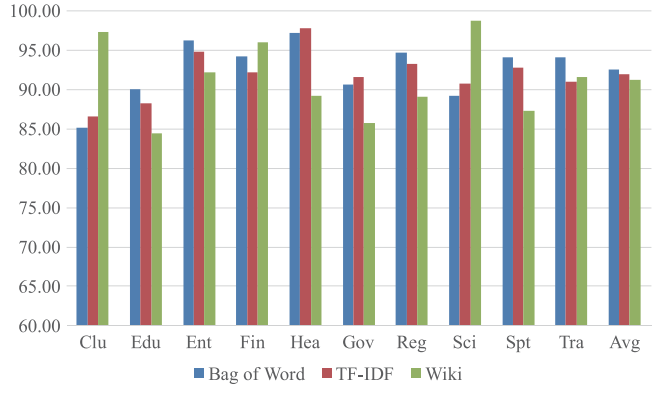
With the explosion of Internet information, the traditional web page classification algorithm based on the training data set model has been unable to handle the complicated web page classification.



(a) Sina.



(b) Ifeng.



(c) Yahoo!.

Fig. 5. Classification results comparing with Wikipedia, BOG and TF–IDF algorithms.

This paper introduced a classification method using WKN. As the most popular scientific knowledge database in the world, Wikipedia knowledge base contains more than 200 types of the linguistic knowledge base. We used Chinese knowledge base to solve the Chinese web page classification, and this approach can be extended to other languages.

At first, we referred to some Chinese portal sites to define PC space. Then, some Wikipedia category vocabularies were defined as EKs in the class space. Several Wikipedia category words with closer distance to keywords were estimated as template keywords in the PA. And then, we estimated the class probability of each WEWs using the RA algorithm. To reflect the affinity of nodes inertial problem, we proposed the inheritance similarity and KRA algorithm. Keywords of web pages were extracted referring to WEWs and the weights were estimated with term frequency and word length. We proposed an inverse sigmoid function to optimize the mapping relation between keyword CPs and page CPs. Experimental results are shown that the system can achieve a good-quality environment oriented to different web page classifications.

The PC definition is very important in our approach, and we will try to give a more flexible approach to reduce the category class drift problem. More discussion about the change of knowledge ontology generation and web page classification will be done in the near future.

Acknowledgments

This work was supported by the NSFC (No. 61502247, 11501302, 61502243), China Postdoctoral Science Foundation (No. 2016M600434), Natural Science Foundation of Jiangsu Province (BK20140895), Scientific and Technological Support Project (Society) of Jiangsu Province (No. BE2016776), and Postdoctoral Science Foundation of Jiangsu Province (1601128B).

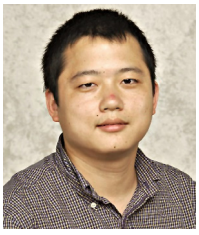
References

[1] K.S. Jones, A statistical interpretation of term specificity and its application in retrieval, J. Doc. 28 (1) (1972) 11–21.
[2] B. Choi, Z. Yao, Web page classification, in: Foundations and Advances in Data Mining, Springer, 2005, pp. 221–274.
[3] X. Qi, B.D. Davison, Web page classification: Features and algorithms, ACM Comput. Surv. 41 (2) (2009) 12.
[4] T. Joachims, Learning to Classify Text Using Support Vector Machines: Methods, Theory and Algorithms, Kluwer Academic Publishers, 2002.
[5] A. Selamat, S. Omatu, Web page feature selection and classification using neural networks, Inform. Sci. 158 (2004) 69–88.
[6] T. Joachims, A probabilistic analysis of the rocchio algorithm with tfidf for text categorization., Tech. rep., DTIC Document, 1996.
[7] L. Jiang, Z. Cai, H. Zhang, D. Wang, Naive bayes text classifiers: a locally weighted learning approach, J. Exp. Theor. Artif. Intell. 25 (2) (2013) 273–286.
[8] M. Rusinol, J. Llados, Logo spotting by a bag-of-words approach for document categorization, in: 10th International Conference on Document Analysis and Recognition, 2009, ICDAR'09, IEEE, 2009, pp. 111–115.
[9] D. Mladenić, Feature subset selection in text-learning, in: European Conference on Machine Learning, Springer, 1998, pp. 95–100.
[10] G. Salton, C. Buckley, Term-weighting approaches in automatic text retrieval, Inf. Process. Manage. 24 (5) (1988) 513–523.
[11] D.D. Jung Y, H. Park, An effect term weighting scheme for information retrieval., Tech. rep., Department of Computer Science, University of Minnesota, 2000.
[12] C. Buckley, G. Salton, J. Allan, The effect of adding relevance information in a relevance feedback environment, in: Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Springer-Verlag New York, Inc., 1994, pp. 292–300.
[13] T. Joachims, Text categorization with support vector machines: Learning with many relevant features, in: European Conference on Machine Learning, Springer, 1998, pp. 137–142.
[14] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I.H. Witten, The weka data mining software: an update, ACM SIGKDD Explor. Newsl. 11 (1) (2009) 10–18.
[15] C.B. Markwardt, Non-linear least squares fitting in idl with mpfit, arXiv preprint arXiv:0902.2850.
[16] J. Zhang, Y. Xiang, Y. Wang, W. Zhou, Y. Xiang, Y. Guan, Network traffic classification using correlation information, IEEE Trans. Parallel Distrib. Syst. 24 (1) (2013) 104–117.
[17] Y. Wang, Y. Xiang, J. Zhang, W. Zhou, G. Wei, L.T. Yang, Internet traffic classification using constrained clustering, IEEE Trans. Parallel Distrib. Syst. 25 (11) (2014) 2932–2943.
[18] J. Zhang, X. Chen, Y. Xiang, W. Zhou, J. Wu, Robust network traffic classification, IEEE/ACM Trans. Netw. 23 (4) (2015) 1257–1270.
[19] P.R. Rijnbeek, J.A. Kors, Finding a short and accurate decision rule in disjunctive normal form by exhaustive search, Mach. Learn. 80 (1) (2010) 33–62.
[20] Y. Zheng, C. Sun, C. Zhu, X. Lan, X. Fu, W. Han, Lwcs: A large-scale web page classification system based on anchor graph hashing, in: 2015 6th IEEE International Conference on Software Engineering and Service Science (ICSESS), IEEE, 2015, pp. 90–94.
[21] H. Wang, F. Nie, H. Huang, Large-scale cross-language web page classification via dual knowledge transfer using fast nonnegative matrix trifactorization, ACM Trans. Knowl. Discov. Data 10 (1) (2015) 1.
[22] S.T. Marath, M. Shepherd, E. Milios, J. Duffy, Large-scale web page classification, in: 2014 47th Hawaii International Conference on System Sciences (HICSS), IEEE, 2014, pp. 1813–1822.

[23] O.R. Zaïane, M.-L. Antonie, Classifying text documents by associating terms with text categories, in: Australian Computer Science Communications, Vol. 24.2, Australian Computer Society, Inc., 2002, pp. 215–222.
[24] G.A. Miller, Wordnet: a lexical database for english, Commun. ACM 38 (11) (1995) 39–41.
[25] C. Matuszek, J. Cabral, M.J. Witbrock, J. DeOliveira, An introduction to the syntax and content of cyc, in: AAAI Spring Symposium: Formalizing and Compiling Background Knowledge and Its Applications to Knowledge Representation and Question Answering, Citeseer, 2006, pp. 44–49.
[26] J. McCarthy, M. Minsky, A. Sloman, L. Gong, T. Lau, L. Morgenstern, E.T. Mueller, D. Riecken, M. Singh, P. Singh, An architecture of diversity for commonsense reasoning, IBM Syst. J. 41 (3) (2002) 530–539.
[27] P. Schönhofen, Identifying document topics using the wikipedia category network, Web Intell. Agent Syst. 7 (2) (2009) 195–207.
[28] P. Wang, J. Hu, H.-J. Zeng, Z. Chen, Using wikipedia knowledge to improve text classification, Knowl. Inf. Syst. 19 (3) (2009) 265–281.
[29] G. Spanakis, G. Siolas, A. Stafylopatis, Exploiting wikipedia knowledge for conceptual hierarchical clustering of documents, Comput. J. 55 (3) (2012) 299–312.
[30] H. Li, G. Sun, B. Xu, L. Li, J. Huang, K. Tanno, W. Wu, C. Xu, An information classification approach based on knowledge network, in: 2014 IEEE 8th International Symposium on Embedded Multicore/Manycore SoCs (MCSoc), IEEE, 2014, pp. 3–8.
[31] Z. Xu, X. Luo, L. Mei, C. Hu, Measuring the semantic discrimination capability of association relations, Concurr. Comput. Pract. Exp. 26 (2) (2014) 380–395.
[32] Z. Xu, X. Luo, S. Zhang, X. Wei, L. Mei, C. Hu, Mining temporal explicit and implicit semantic relations between entities using web search engines, Future Gener. Comput. Syst. 37 (2014) 468–477.
[33] Z. Xu, Y. Liu, L. Mei, C. Hu, L. Chen, Generating temporal semantic context of concepts using web search engines, J. Netw. Comput. Appl. 43 (2014) 42–55.
[34] M. Völkel, M. Krötzsch, D. Vrandecic, H. Haller, R. Studer, Semantic wikipedia, in: Proceedings of the 15th International Conference on World Wide Web, ACM, 2006, pp. 585–594.
[35] S. Adafre, M. de Rijke, Discovering missing links in wikipedia, in: Proceedings of the 3rd International Workshop on Link Discovery, ACM, 2005, pp. 90–97.
[36] X. Hu, X. Zhang, C. Lu, E. Park, X. Zhou, Exploiting wikipedia as external knowledge for document clustering, in: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2009, pp. 389–396.
[37] J. Min, G. Jones, Building user interest profiles from wikipedia clusters, in: Proceedings of the Workshop on Enriching Information Retrieval, ENIR 2011, at SIGIR 2011.
[38] M. Charikar, Similarity estimation techniques from rounding algorithms, in: Proceedings of the Thiry-fourth Annual ACM Symposium on Theory of Computing, ACM, 2002, pp. 380–388.
[39] W.J. Heeringa, Measuring dialect pronunciation differences using levenshtein distance (Ph.D. thesis), Citeseer, 2004.
[40] H. Li, X. Xu, L. Lai, Y. Shen, Online commercial intention detection framework based on web pages, Int. J. Comput. Sci. Eng. 12 (2/3) (2016) 176–185.
[41] S. Cesare, Y. Xiang, W. Zhou, Malwise: an effective and efficient classification system for packed and polymorphic malware, IEEE Trans. Comput. 62 (6) (2013) 1193–1206.

Huakang Li was born in Suzhou, China. He received the Master and Ph.D. degree from the School of computer science and engineering, University of Aizu, Aizuwakamatu, Japan, in 2007 and 2011, respectively. He is currently working in the School of Computer Science and Technology, School of Software, Nanjing University of Posts and Telecommunications. His current research interests include big data mining, web mining, social network, user profile and semantic Web [31–33]. He has authored or co-authored more than 20 publication include IEEE and ACM. He has served as a program chair and TPC member for several international conferences and journal like TPDS, HRI, FCST, NBiS, RACS, AINA, CSE. He is a member of IEEE CS, and ACM, and CCF China.

Zheng Xu was born in Shanghai, China. He received the Diploma and Ph.D. degrees from the School of Computing Engineering and Science, Shanghai University, Shanghai, in 2007 and 2012, respectively. He is currently working in the third research institute of ministry of public security and the postdoctoral in Tsinghua University, China. His current research interests include topic detection and tracking, semantic Web and Web mining. He has authored or co-authored more than 70 publications including IEEE Trans. On Fuzzy Systems, IEEE Trans. On Automation Science and Engineering, IEEE Trans. On Cloud Computing, IEEE Trans. On Emerging Topics in Computing, IEEE Transactions on Systems, Man, and Cybernetics: Systems, etc.



Tao Li is currently a full professor in the School of Computer Science, Florida International University, and the head of School of Computer Science and Technology, School of Software in Nanjing University of Posts and Telecommunications, China. He received his Ph.D. in computer science from the Department of Computer Science, University of Rochester in 2004 (My old homepage at Rochester). He was a recipient of NSF CAREER Award (2006–2010) and multiple IBM Faculty Research Awards (2005, 2007 and 2008). In 2009, he received FIU's Excellence in Research and Creativities Award. In 2010, he received IBM Scalable Data Analytics Innovation Award. He received the inaugural Graduate Student Mentorship Award from the College of Engineering and Computing at FIU in 2011. He is on the editorial board of ACM Transactions on Knowledge Discovery from Data (ACM TKDD), IEEE Transactions on Knowledge and Data Engineering (IEEE TKDE), and Knowledge and Information System Journal (KAIS).

Guozi Sun is currently a professor of the School of Computer, Nanjing University of Posts and Telecommunications, China. His recent research interests mainly include digital forensics, multimedia forensics, social network forensics, and digital investigation. Dr. Sun has published more than 100 refereed papers in the academic journals and international conference proceedings in the related research areas. He has served as a program chair and TPC member for several international conferences, and editor-in-chief, associate editor, editorial board member and guest editor for a number of scientific journals. He is a member of IEEE CS, and ACM, USA, and CCF, CIE, ISFS, China.

Kim-Kwang Raymond Choo received the Ph.D. degree in Information Security from Queensland University of Technology, Australia, in 2006. He currently holds the Cloud Technology Endowed Professorship at The University of Texas at San Antonio. He serves as the Special Issue Guest Editor of ACM Transactions on Embedded Computing Systems (2017), ACM Transactions on Internet Technology (2016), Digital Investigation (2016), Future Generation Computer Systems (2016), IEEE Cloud Computing (2015), IEEE Network (2016), Journal of Computer and System Sciences (2017), Multimedia Tools and Applications (2017), Pervasive and Mobile Computing (2016), etc. He is a recipient of various awards, including the ESORICS 2015 Best Paper Award, Winning Team of the Germany's University of Erlangen-Nuremberg (FAU) Digital Forensics Research Challenge 2015, and the 2014 Highly Commended Award by the Australia New Zealand Policing Advisory Agency, the Fulbright Scholarship in 2009, the 2008 Australia Day Achievement Medallion, and the British Computer Society's Wilkes Award in 2008. He is a Fellow of the Australian Computer Society, and a Senior Member of IEEE.