International Journal of Hybrid Information Technology
Vol. 9, No.8 (2016), pp. 223-232
http://dx.doi.org/10.14257/ijhit.2016.9.8.20

Matching Mechanism of Web Service Based on Hybrid Semantic Similarity Computation

Peng Li[1]

1. School of Information Science and Engineering, Hunan First Normal University, Changsha, China
goodbetter@163.com

Abstract

Fuzzy search algorithm with hybrid semantic similarity is proposed aiming at serious performance issues in information retrieval, which results from fuzzy search condition, in search engine land. First of all, new conceptual extraction method is proposed according to similarity calculation concept; after that, TF-IQF is adopted to divide the link into tabs, and the set comprised of these tabs is used for indicating the query; ultimately, binary graph is constructed to identify related queries and is employed to compute query similarity. Experimental results show that proposed algorithm has achieved better recall ratio, retrieval precision and F-measure compared with clicking document, related queries and reverse query algorithm.

Keywords*: Fuzzy query; Similarity calculation; Binary graph; Web service*

1. Introduction

Nowadays, the webpage recorded by search engine [1] is increased rapidly as sustained expanding of internet. As plain as the nose on your face, new information will be generated rapidly when people getting a large number of useful information from internet, thus causing major issues [2] of low signal-to-noise ratio (ratio of useful information found and all information) and poor efficiency of dealing with different type of information. This is not because if information content is enough or not, in contrast, but because it's oversized with different types of formats, and not all information are valuable, the result gives rise to overloading information for users.

Based on simple search query [3], it has become more and more difficult to find related information which meets the demand of users in modern information environment with such massive data size, because keywords submitted to search engine by users generally are short and fuzzy. The research found that the average query length submitted to search engine is only 2.35 words. The researchers found that the average query length is 1.8 words according to user log analysis [5] of Sogou search engine in China, at the same time, there are approximately 93.15% users whom the number of word query is less than 3. Obviously, these short and fuzzy query words are impossible to accurately express what users really need, the users will certainly obtain a great deal of webpage information, which is irrelevant to self-demand, according to such fuzzy search [6]. In addition, users might not adopt more search terms to rewrite the query, for an additional burden will occur as they are searching.

The main reason why a user clicks some search result is such webpage contains relevant themes which interest them, hybrid semantic strategy of evaluating query similarity is proposed based on clicking data of verifying the content that interests users. Three methods are adopted to calculate query similarity in this article, namely word similarity [7], conceptual extraction [8] and TF-IQF model [9] of VSM (vector space modal). Hybrid semantic similarity strategy proposed mainly the result of extensive discussion to two latter methods. Such method chiefly contains three steps below: (1)

ISSN: 1738-9968 IJHIT

after submitting the search by users, query concept (namely important vocabulary and phrases of such category in network segment) or tabs (significant terms in URL clicked) and their mutual relationship are digged out from network segment to structure binary graph; (2) query similarity is calculated based on above binary graph, in the meantime, calculation method of hybrid similarity is proposed; (3) the most similar search is recommended to users to simplify the search.

2. Related Work

In network search engine, clicking data [10] is a kind of implicit feedback of user information. Apparently, it's an important resource of query suggestion [11]. Beeferman and Berger proposed a kind of clustering algorithm [12], user query log is adopted to cluster URLs and queries to find out related query. Binary graph is adopted by them as shown in Figure 1. Left node in Figure 1 represents query, and the left one represents URLs clicked by users. If the user clicks one URL, related query and URL of it will be structured on binary chart. After obtaining binary chart, iterative algorithm is adopted to cluster two queries and two URLs. Disadvantage of such algorithm is that it cannot effectively handle noisy data, that is, if the user click one URL mistakenly, then two irrelevant queries will be connected with one another for good.

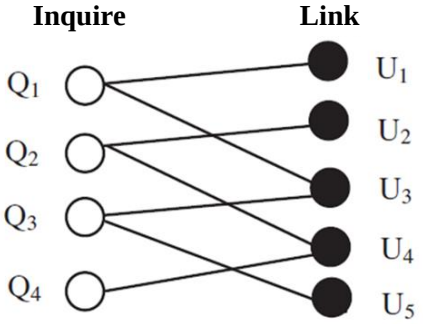

Figure 1. Query-Link Binary Chart

To calculate similarity among different queries, the similarity of documents is also taken into consideration in literature [13]. They proposed that if same or similar words are contained in two queries, then they should be clustered together, thus resulting in selecting the same URLs. However, there are only so few URLs that point to same document because the query condition is generally too short, their method is not effective to eliminate ambiguity of Web query. Moreover, such method is required to pre-built document classification system with higher classification precision, for its complex calculation.

In summary, the two methods have a common major problem, there are very few URLs quantity of common click between different queries, and there are a handful of popular queries are provided with enough information to dig out URLs they click together. Therefore, it's a chance in a million that the users would see the same query results, not to mention to click them. To resolve such problem, Leung et.al proposed a viewpoint based on conceptual graphs [14], they consider extracting concept from network segment and Beeferman and Berger method is adopted to such new text. The volume of result documents is reduced with the use of concept, at the same time, it maintains an precision and capacity to meet the demand of users. Literature [15] proposed query algorithm, which is different from others, with the use of TF-IQF model. URL character string is divided into some tabbing vocabularies according to some separate tabs , and the weight of these tabs are also measured to calculate their similarity.

3. Hybrid Semantic Similarity Algorithm

3.1. Clicking Document

Implicit feedback and query mode can be used to organize network file, namely the query of users can be seen as documents characteristic vocabulary to resolve dictionary problem. Oppositely, the users may select graph document as an extension of query vocabulary. Doc1 and Doc2 can be used as description for Query 1, Doc2 and Doc4 can be used as description for Query 3. Therefore, such method may handle issues of being lack of query specification. In other words, if URLs is clicked during query search, the graph documents clicked corresponding to these URLs can be regarded as similar or related documents.

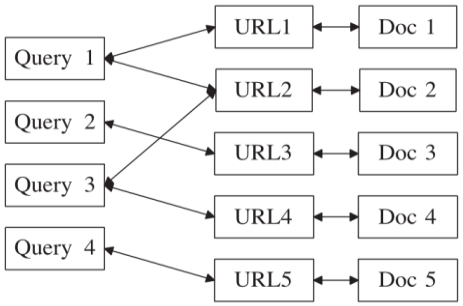

Figure 2. Clicking Document

Assume that two queries are p and q, a $m \times m$ matrix $S(s_{ij})_{m \times m}$ is obtained, which shows similarity relation of documents.

$$S(s_{ij})_{m \times m}^{p,q} = \begin{bmatrix} s_{11} & s_{12} \ldots s_{1m} \\ s_{21} & s_{22} \ldots s_{2m} \\ \ldots & \ldots \ldots \ldots \\ s_{m1} & s_{m2} \ldots s_{mm} \end{bmatrix} \quad (1)$$

Where, $s_{ij} = sim(d_i, d_j)$, $i = 1,2,...,m$, $j = 1,2,...,m$, d_i and d_j represent two clicking documents, m represents total number of clicking document which is related to p and q. Obviously, that is:

(1) $s_{ij} = 1, (if \quad d_i = d_j \neq \varnothing, i = 1,2,...,m; j = 1,2,..,m;)$

(2) $s_{ij} = 0, (if \quad d_i \neq \varnothing \ or \ d_j = \varnothing, i = 1,2,...,m; j = 1,2,..,m;)$

$sim(d_i, d_j)$ can be measured with cosine algorithm, namely the inner product of normalization of these two correlation matrixs.

$$sim(d_i, d_j) = \cos\theta = \frac{V(d_i) \cdot V(d_j)}{|V(d_i) \times V(d_j)|} = \frac{\sum_{k=1}^{n} w_{i,k} \times w_{j,k}}{\sqrt{\sum_{k=1}^{n} w_{i,k}^2 \sum_{k=1}^{n} w_{j,k}^2}} \quad (2)$$

Where, $w_{i,k}$ is associated with vocabulary t_i in document d_k, it's calculated through $t_f \times idf$ method. Therefore, measurement $sim_{doc}(p,q)$ definition of clicking document is used for searching similarity of p and q

$$sim_{doc}(p,q) = \frac{1}{2} * \left(\frac{\sum_{i=1}^{m} \max(s_{i1}, s_{i2}, ..., s_{im})}{u_n(p)} + \frac{\sum_{j=1}^{m} \max(s_{1j}, s_{2j}, ..., s_{mj})}{u_n(q)} \right) \quad (3)$$

Where, s_{ij} is an element of matrix $S(s_{ij})_{m \times m}$, $u_n(p)$ and $u_n(q)$ are total number of query p and q corresponding to query documents of URLs clicked.

3.2. Relation Query

As we know, the similarity of these two queries is higher if there appear to be more synonyms. If some original query has same or similar vocabulary with other query, then they are might be coordinate indexing. In other words, the quality of query expansion can be improved if to retrieve similar query. Therefore, relation query can be retrieved as expanded vocabulary of candidate query. If the format of clicking through data by users is $Click-Through_i = (userid_i, query_i, clicked\_url_i)$, process of original query is transformed into $query = (q_1, q_2, ...q_3)$. For example, input query West Lake can be denoted as $(west, lake)$. If $(q_1 \in query_i) \wedge (q_2 \in query_i) \wedge ... \wedge (q_n \in query_i)$, $i = 1, 2, ..., n$, of which n quantity of relation query. List of relation query can be defined as

$$AssociatedQueryList = (assQ_1, assQ_2, assQ_3, ..., assQ_n) \quad (4)$$

Therefore, similarity of relation query is defined as

$$sim_{ass}(p,q) = sim_{cn}(p,q) + \delta \cdot sim_{sn}(p,q) \quad (5)$$

Where, δ is an real constant between 0 and 1. $sim_{cn}(p,q)$ represents similarity based on lexical item of literal words, it's defined as

$$sim_{cn}(p,q) = \frac{t_{cn}(p,q)}{\max(t_n(p), t_n(q))} \quad (6)$$

Where, $t_n(p)$ and $t_n(q)$ are intersected word numbers of query p and q , respectively. $t_{cn}(p,q)$ is quantity of same vocabulary in p and q . For example, $sim_{cn}(p,q)$ of query "compute game'' and ''mobile game'' is $sim_{cn}(p,q) = 1/\max(2,2) = 0.5$. $sim_{sn}(p,q)$ is used for representing semantic similarity of query p and q . Here query-concept binary graph is adopted to calculate semantic similarity between query, see the formula (5).

3.3. Reverse Query

Two queries corresponding to same URL are highly correlated when URL is clicking. For example, because two queries such as "cross fire" and "world of war craft" are given similar URL, games.sina.com.cn, they are semantic relativity. Such two queries can be seen as reverse query because they all want to search games. Similarly, queries like "audi" and "ford" are as well semantic relativity, for their same URL car.auto.ifeng.com shows that they all want to search vehicles. Figure 3 shows that an example of reverse query, that is, $Query_1$ and $Query_3$ have semantic relativity with $Query_4$ and $Query_5$, respectively.

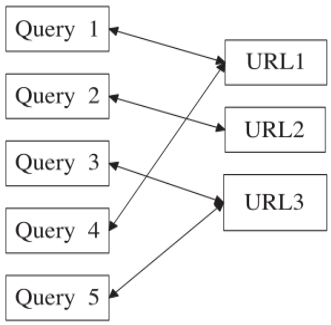

Figure 3. Reverse Query

It's verified that the performance resulted from calculating semantic similarity between queries using query -(TF-IQF) binary graph is obviously superior to the ones obtained using URLs alone.

3.4. Hybrid Semantic Similarity Algorithm

In general, each above-mentioned method may partly uncover limited sematic information what is hidden in query definition. These methods, however, still have the flaw and it's insufficient of capturing deep-seated sematic information. A new method that integrated these three strategies is proposed, it's defined as

$$sim(p,q) = \alpha * sim_{doc}(p,q) + \beta * sim_{ass}(p,q) + \gamma * sim_{rev}(p,q) \quad (7)$$

Where, $sim_{doc}(p,q)$ is similarity of document clicked, $sim_{ass}(p,q)$ is similarity of relation query, $sim_{rev}(p,q)$ is similarity of reverse query. α, β and γ are real constant between 0 and 1 and satisfy restricting $\alpha + \beta + \gamma = 1$. To find out optimal weight of α, β and γ, control experiment will be conducted by adjusting weight step by step in experimental section.

3.5. Clustering

Clustering algorithm based on GA (genetic algorithm) is adopted to query in group and generate subtitle structure for users, algorithm description is:

(1) First, each query is viewed as individual point in query space, then integrate query randomly and write in string form, it's called as chromosome. The set of chromosome is called as species, then an randomly distributed species is created.

(2) Use three species to enlighten operator, namely: selection, intersection and mutation are applied to process of generating new sub-chromosome. These three operations will be continued for several generations until it meets the final criterion. In this search, robust chromosome with high fitness is selected to produce the next generation so as to keep good seeds. Classic single-point crossover and Gaussian mutation are adopted in this method.

(3) The algorithm won't be terminated until when non-enhancing optimal chromosome continuously produces iteration for $n_{\max}(n_{\max} = 10)$ generations. In an iteration process, formula (8) is adopted to calculate the similarity of two clustering.

$$sim(c_i, c_j) = \frac{\sum_{i=1}^{n_1} \sum_{j=2}^{2} sim(q_i, q_j)}{n_1 \times n_2} \quad (8)$$

Where, q_i and q_j are queries in clustering c_i and c_j, respectively, n_1 and n_2 are query numbers in clustering c_i and c_j.

4. Experiment

To evaluate the performance of algorithm proposed, experimental procedure of collecting clicking data needed will be first described. Use Google to search 300 query conditions given and collect clicking data. To avoid an deviation, query tested is randomly selected from 10 kinds of different categories. Table 1 shows 10 theme categories of query selected. Top 50 search results (network segment) queried are collected from Google as data corpus.

Table 1. Test 10 Theme Categories of Query

Theme	Description
1	Journey
2	Computer
3	Car
4	Fruit
5	Game
6	Education
7	Military
8	Health
9	Digital Product
10	Sport

To compare algorithm proposed with four algorithms like clicking document, correlation query, reverse query and hybrid semantic similarity, index like recall ratio, precision and F-measure are evaluated in the experiment. Assume that query q exists, query category generated from clustering algorithm is $\{q_1, q_2, ...\}$, recall ratio $\{q_1, q_2, ...\}$, precision $\mathrm{P}(q)$ and F-measure are calculated with algorithm below

$$R(q) = \frac{|q\_relevant \cap q\_retrieved|}{|q\_relevant|} \tag{9}$$

$$P(q) = \frac{|q\_relevant \cap q\_retrieved|}{|q\_retrieved|} \tag{10}$$

$$F - Measure = 2 \times \frac{R(q) \times P(q)}{R(q) + P(q)} \tag{11}$$

Of which $q\_relevant$ query set existed in predefined clustering of query q. $q\_retrieved$ is relevant query set $\{q_1, q_2, ...\}$ generated from algorithm proposed. Compared with recall ratio, precision and F-measure index of these four methods. Where, α, β and γ are three weight parameters in formula (7). These three value first are altered with 0.1 interval in the experiment. sim_{doc}, sim_{ass} and sim_{rev} represent clicking document, correlation query and reverse query method, respectively. When $\alpha = 0.1, \beta = 0.6, \gamma = 0.3$, F-measure performance is the highest that far higher than other three methods, and also higher than the performance of hybrid semantic similarity algorithm when α, β and γ parameters are other value. For the result of recall ratio, although the performance is optimal when $\alpha = 0.3, \beta = 0.5, \gamma = 0.2$, the precision and F-measure both are lower. At the same time, when $\alpha = 0.4, \beta = 0.5, \gamma = 0.1$, the precision of it is the highest. When $\alpha = 0.1, \beta = 0.6, \gamma = 0.3$, F-measure performance of hybrid semantic similarity algorithm is optimal. By adjusting these three parameters, F-measure performance of such algorithm is optimal when $\alpha = 0.15, \beta = 0.55, \gamma = 0.3$, namely β and γ value are greater than α value, this is benefited from comparing sim_{ass}, sim_{rev} and sim_{doc} methods, the latter two methods may provide more semantic information and descriptive expression for query, however, simply clicking document method is only applied to independent vocabulary. Figure 4 and Figure 5 represent precision of similarity and description comparison of four methods, Figure 6 shows comparison of precision and recall ratio.

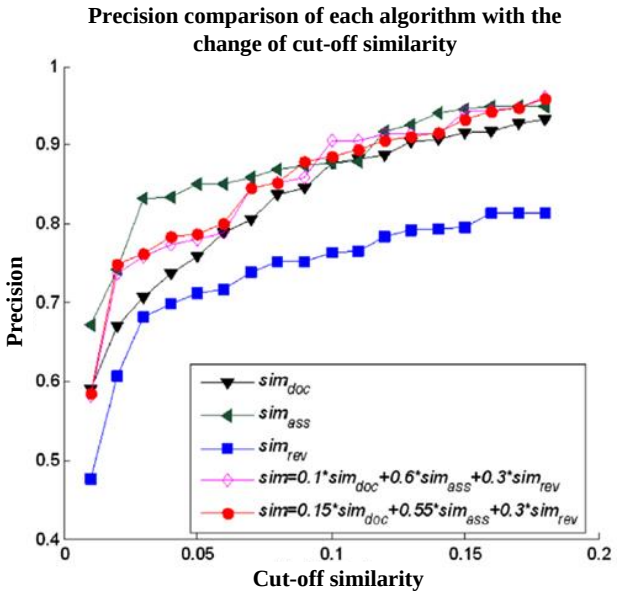

Figure 4. The Precision of Each Algorithm Changing with the Similarity Comparison

sim_{rev} method's precision is the lowest as shown in Figure 4, sim_{ass} has the highest precision but recall ratio goes the other way as shown in Figure 5. In addition, Figure 5 shows that sim_{doc} 's recall ratio being the lowest, however, hybrid semantic similarity algorithm proposed has the highest recall ratio. Therefore, you can easily see hybrid semantic similarity algorithm proposed both may obtain better recall ratio and precision as shown in Figure 6.

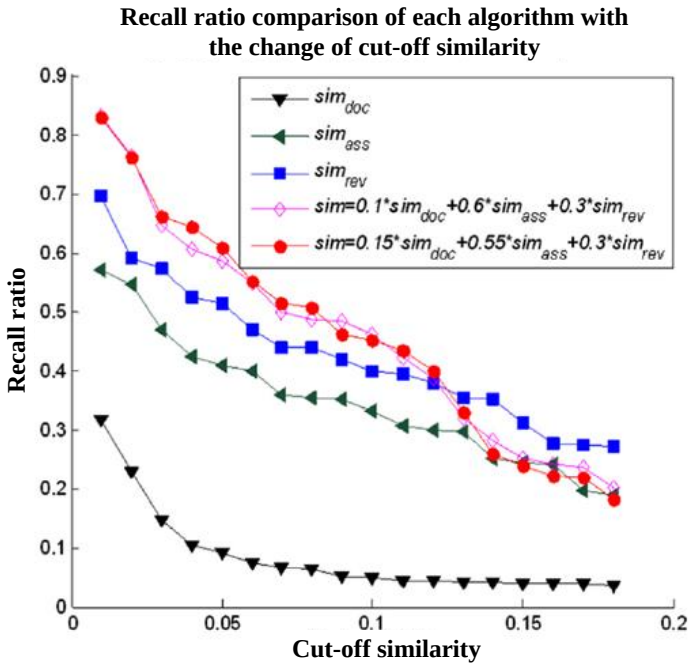

Figure 5. Recall Ratio Comparison of Each Algorithm by Changing Similarity

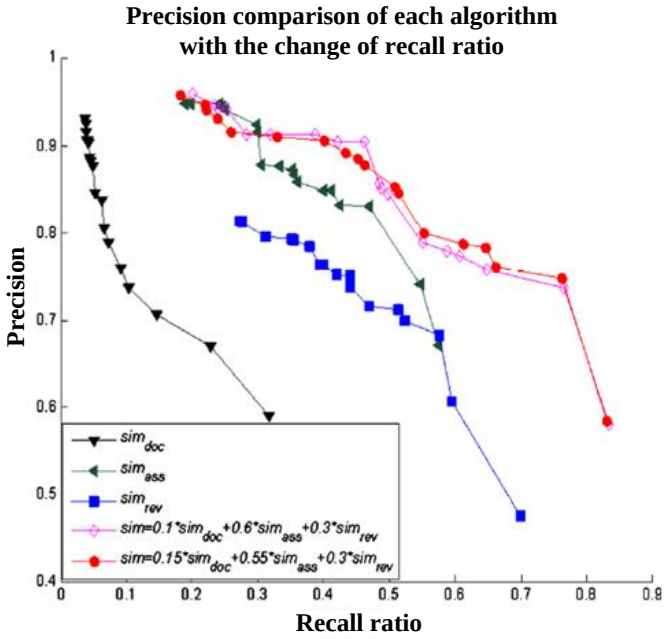

Figure 6. Comparison of Precision and Recall

5. Conclusion

Hybrid semantic similarity method is proposed aiming at major problem of fuzzy search condition in search engine land. First, new conceptual extraction method which is conceptually close to other method is proposed. After that, TF-IQF model is adopted to divide URLs into tabs, hence query can be represented by the set comprised of these tabs. Ultimately, binary graph is used for calculating query similarity. To calculate query similarity in semantic analysis, three kinds of query methods are given, that is: clicking document, correlation query and reverse query. Hybrid semantic method is proposed by setting up an appropriate weight for three methods and some control experiments are conducted. The results show hybrid semantic method proposed has obtained a high performance assessment. In the future, various weight setting will be varied, the effect of different weight variations on query will be investigated and efficiency of query will be further improved to meet diversified demand of users.

Acknowledgement

Search Is supported by a science and technology plan project of Hunan province (No. 2014GK3018); a project supported by scientific research fund of Hunan provincial education department (No. 12C0593) and a project of the key laboratory of basic education informatization technology in Hunan province (No. 2015TP1017)

References

[1] Y. Liang*, "Correlations between Health-Related Quality of Life and Interpersonal Trust: Comparisons Between Two Generations of Chinese Rural-to-Urban Migrants", Social Indicators Research, vol. 123, no. 3, pp. 677-700.

[2] Y. Liang* and P. Lu, "Medical insurance policy organized by Chinese government and the health inequity of the elderly: longitudinal comparison based on effect of New Cooperative Medical Scheme on health of rural elderly in 22 provinces and cities", International Journal for Equity in Health, 13:37, 1-15. DOI:10.1186/1475 -9276-13-37, **(2014)**.

[3] Y. Liang*, D. Zhu, "Subjective Well-Being of Chinese Landless Peasants in Relatively Developed Regions: Measurement Using PANAS and SWLS". Social Indicators Research, vol. 123, no. 3, pp. 817-835.
[4] Y. Liang* and X. Wang, "Developing a new perspective to study the health of survivors of Sichuan earthquakes in China: a study on the effect of post-earthquake rescue policies on survivors' health-related quality of life", Health Research Policy and Systems, vol. 11, pp. 41, 1-12. DOI:10.1186/1478-4505-11-41.
[5] J. Hu and Z. Gao, "Modules identification in gene positive networks of hepatocellular carcinoma using Pearson agglomerative method and Pearson cohesion coupling modularity", Journal of Applied Mathematics, **(2012)**.
[6] Y. Geng, J. Chen, R. Fu, G. Bao and K. Pahlavan, "Enlighten Wearable Physiological Monitoring systems: On-Body RF Characteristics Based Human Motion Classification Using a Support Vector Machine", vol. 99, **(2015)**, pp. 1-16.
[7] X. Song and Y.Geng, "Distributed Community Detection Optimization Algorithm for Complex Networks", Journal of Networks, vol. 9, no. 10, **(2014)**, pp.2758-2765.
[8] K. Pahlavan, P. Krishnamurthy and Y. Geng, "Localization Challenges for the Emergence of the Smart World", Access, IEEE, vol. 3, no. 1, **(2015)**, pp.1-11.
[9] J. He, Y. Geng, Y. Wan, S. Li and K. Pahlavan, "A cyber physical test-bed for virtualization of RF access environment for body sensor network", Sensors Journal, IEEE, vol. 13, no.10, **(2013)**, pp. 3826-3836.
[10]Z. Lv, A. Tek and F. Da Silva, "Game on, science-how video game technology may help biologists tackle visualization challenges", PloS one, vol.8, no. 3, **(2013)**, pp.57990.
[11] T. Su, W. Wang and Z. Lv, "Rapid Delaunay triangulation for randomly distributed point cloud data using adaptive Hilbert curve", Computers & Graphics, vol. 54, **(2016)**, pp. 65-74.
[12] J. Hu, Z. Gao and W. Pan, "Multiangle Social Network Recommendation Algorithms and Similarity Network Evaluation", Journal of Applied Mathematics, **(2013)**.

Author

Peng Li, He received his B.E. degree from Hunan University in 1996, and M.S. degree from Hunan University in 2013. He is currently a lecturer in School of Information Science and Engineering, Hunan First Normal University, China. His research interest involves Web service, the semantic Web, computer network, and information security.

