Extracting Subject from Internet News by String Match*

YIN Zhong-hang, WANG Yong-cheng, CAI Wei, HAN Ke-song

(*School of Electronics and Information Technology, Shanghai Jiaotong University, Shanghai* 200030, *China*)
E-mail: liaoning-yin@263.net
http://www.sjtu.edu.cn
Received December 21, 2000; accepted July 12, 2001

Abstract: Subject extraction from a text is very important for natural language processing. Traditional methods mainly depend on the mode of "thesaurus plus match". It is not fit to process Internet news because of its limited volume and slow update speed. After analyzing the news structure carefully, this paper presents a new practical method to extract news subjects without thesaurus, and give the main implementing procedure. Instead of large thesaurus, it uses the special structure of Internet news to find the repeated strings. These repeated strings could express the news subjects very well. Experimental results show that this method can extract the most important subject strings from most of Internet news rapidly and efficiently. Moreover, this method is equally efficient to other Asian languages such as Japanese and Korean, as well as other western languages.
Key words: Web information processing; Internet news; subject extraction; string match; thesaurus

Internet news plays a very important role in the huge Internet information set. According to the latest statistic report of CNNIC, 84.38% of information got from Internet by Chinese users is news[1]. Therefore, research on Internet news is becoming a hotspot in NLP field. One of the challenges we meet is subject extraction from Internet news. Traditional methods of subject extraction from a text mainly depend on the mode of "thesaurus plus match"[2]. Two problems arise when processing Internet news with thesaurus. One is the limited volume of thesaurus, compared with the unlimited topic scope of Internet news. The other is the slow speed of new concept update manually, compared with the uninterrupted emergence of new concepts on Internet nearly all the time. Normally, these new concepts, such as people names or organization names, play an important role in NLP. Therefore, the performance of news subject extraction with traditional thesaurus is not ideal. For instance, the subject precision of SE (Subject Explore) system, which collected 163 599 concepts in its thesaurus, only reaches 81% when processing Internet news. After analyzing subject elements and studying the structure of Internet news, we present a new idea of extracting subject directly from Internet news without thesaurus or word segmentation. Exploiting the special structure of Internet news, we search the full body with title characters to find the repeated strings. These repeated strings would express the news subjects very well after simple processing. Our experiments show that this method can efficiently process nearly all of Internet news. The precision is 95.10% for all extracted strings, and the ratio up

* Supported by the National Natural Science Foundation of China under Grant No.60082003 (国家自然科学基金)

YIN Zhong-hang was born in 1968. He is a Ph.D. candidate at School of Electronics and Information Technology, Shanghai Jiaotong University. His current research interests include Internet information processing and text data mining. **WANG Yong-cheng** was born in 1939. He is a professor and doctoral supervisor of the Department of Computer Science and Engineering, Shanghai Jiaotong University. His current research areas include natural language processing and artificial intelligence. **CAI Wei** was born in 1970. He is a Ph.D. candidate at School of Electronics and Information Technology, Shanghai Jiaotong University. His current research interests include Web information processing and text data mining. **HAN Ke-song** was born in 1973. He is a Ph.D. candidate at School of Electronics and Information Technology, Shanghai Jiaotong University. His current research interests include Web information processing and text indexing.

© 中国科学院软件研究所 http://www.jos.org.cn

to 97.50% for the former 3 strings. In addition, this method is equally efficient to other Asian languages, such as Japanese and Korean. After simple adjustment, it can also fit for other western languages.

1 News Subject and Structure

1.1 Subject definition

Text subject means the main topic and central idea of a text. It is important for people to communicate with and understand each other. Meanwhile, it is the basic unit for computer to process natural languages[3].

Text subject may be roughly divided into four sorts according to the form of expression:

Subject word. One word (also named as keyword). It can express text subject simply. It is often normalized to construct thesaurus. It is the original form of subject used by computer to process a text.

Subject concept. One or a series of words (phrase). It can express text subject more concretely than single word. It may also be a multiplex concept consisting of several simple concepts.

Subject sentence. A natural sentence. It can express text subject. Text title or subtitle is one kind of subject sentence.

Subject paragraph. A shorter text. It can express the subject of a longer text clearly. Abstract is one kind of often-used subject paragraph.

The relationship among the four sorts is the later including the former. In other words, the longer includes the shorter. For example, a subject paragraph normally includes one or several subject sentences. A subject sentence includes one or several subject concepts, and so on.

The difference between human and computer in expressing text subject is that the former often uses subject sentence or subject paragraph, while the later usually uses subject words because they are easy to assemble. But with the development of NLP, people have found that subject concepts, which have the similar virtue being easy to assemble, can express more concrete and more accurate meanings than subject words in computer. So more and more NLP tasks tend to use subject concepts as the basic unit. For example, they may be used as the basic semantic units of retrieval, summarization, classification and filtering.

1.2 News elements

There are hundreds of books talking about “news” and “news elements”. No matter how old or how new those books are, no matter what philosophy and theories they stand for, all the authors share the same proposition of news as E.L. Shuman’s. According to Frank L.Mott, “perhaps it was Shuman” who included the “Rule of the Five Ws” in his treatise of 1903. Even after nearly 100 years, Internet news, the new kind of news, is not exceptive. The “5 Ws” are “who”, “what”, “where”, “when”, and “why”. Then, “how” is added. In other words, each news should include “6 Ws”: “who”, “what”, “where”, “when”, “why” and “how ” (Some people also call these elements as “5 Ws +1 H”). “6 Ws” are necessary conditions of news[4]. Table 1 shows their function.

Table 1 News elements and distribution

Index	Items	Function	In title	In body	Ratio of subject repeat(%)
1	who	Object (person, organization, etc.) of a news event	497	500	99.40
2	what	Happened, happening things	498	500	99.60
3	where	Place, location	431	450	95.80
4	when	Time things happened	113	490	22.60
5	why	Cause, consequence	15	210	7.140
6	how	Situation and procedure	2	110	1.80

© 中国科学院软件研究所 http://www.jos.org.cn

Further study tells us that “6 Ws” are often expressed in the form of subject concepts, not limited to subject words. In addition, among the six elements, people pay more attention to “who”, “what” and “where”. Other two elements, “why” and “how”, are not the main subjects of news, at least, they are less important than the former three for readers’ first looking. As for the element “when”, people often omit it because it is often means “today” or “now”. In fact, the former three factors are also the most important in NLP. Therefore, our main goal is to extract the former three elements.

1.3 News structure

A news text consists of a title and a body. The two parts mainly serve as following functions:

(1) Title: Defining the scope of the news and the most important details, such as object (who), event (what), or place (where). It is characterized as being explicit and condensed to attract readers’ attention immediately. Otherwise, no one will pay attention to the following content. Its length is limited, from 10 to 20 Chinese characters about. Normally, concepts in title are related to the news subjects. More than 98% of titles can express the news subject clearly[5].

(2) Body: Explaining the news object, event and place (who, what, and where), which have only been simply presented in title because of its limited length. Of course, body also describes other three elements in detail.

We named the reappearance of some important elements between news title and body as subject repeat. Subject repeat offers us a hint: we may obtain many important elements such as who, what, where and when by simple string match. These elements are useful for Internet news processing. Thus we can avoid using thesaurus and compensate for its shortage.

To confirm this thought, we manually checked the distribution of “6 Ws” for the 500 Chinese news texts (2000/5), downloaded randomly from the web of Hong Kong’s InfoBank, the biggest Chinese information base. Table 1 shows our result. It tells us that element “who” in 497 news (99.40%) appearing in titles and bodies at the same time. As for what and where, the two ratios are 99.60% and 95.80%. There are also other elements appearing in titles, but their ratio is limited. From analysis and experiment above, we can reasonably begin to study how to extract important subjects by making use of the phenomena of subject repeat.

2 Implementation

2.1 Subject gene

Subject repeat is the basis of our algorithm. To find the repeated subject strings, we make use of the concept of subject gene as Definition 1.

Definition 1. Subject gene (*SG*) is two or more uninterrupted characters appearing in both title and body at the same time. Two neighbored *SG*s can be connected to form a longer *SG*. As the basic unit of subject strings, it may be a complete concept, or merely a part of concept.

The Purpose we use *SG*s is to connect all possible neighbored *SG*s to form subject strings. Let us illustrate the definition and usage of *SG* by an example. Here is a piece of news:

Example. A piece of news

Title: 合和实业筹建衡阳高速公路

Body: 合和实业计划在湖南省衡阳市以南兴建长达三百多公里的高速公路……

In the news, Chinese character “合” and “和 ” are uninterrupted in both title and body at the same time. So they construct a *SG* “合和”. Equally, “和” and “实” also construct another *SG* “和实”. Because the two *SG*s, “合和” and “和实”, are neighbored, they can be connected to form a longer *SG* “合和实” by deleting the repeated character

© 中国科学院软件研究所 http://www.jos.org.cn

"和". After connecting all possible neighbored *SG*s step by step, we can obtain the final subject strings such as "合和实业(Hehe Industrial)".

2.2 Data structure

The basic data structure is linear list, showed in Figs. 1 and 2. In the two figures, *m* is the number of characters in title, *n* is the number of characters in body, and *s* is the number of original *SG*s. Figure 1 is one of records for a title character, where the number of records is less than or equal to *m* and there are no more than *n*+2 items for each record. It records all the positions of title characters in body. By checking the list, we can find all possible original *SG*s. Figure 2 is one of *SG* records, where *Hpt*, *Hpb*, *Tpt* and *Tpb* stand for the position of each *SG* in title and body. We can know whether two *SG*s are neighbored or not by checking the four items (the neighbor of two *SG*s means one's head is another's tail in both title and body). As for *Len* and *Freq*, we use them to compute the *SG*'s weight.

Cha(*m*)	*Times*	*Pos*1	*Pos*2	…	*Posn*

Cha(*i*): Characters in Title (*i*=1,2,…,*m*)
Times: *Cha*(*i*) appearing times in body
Pos: *Cha*(*i*) different position in body

Fig. 1 Title character linear list

SG(*s*)	*Len*	*Freq*	*Hpt*	*Hpb*	*Tpt*	*Tpb*

Hpt: Head position in title
Hpb: Head position in body
Tpt: Tail position in title
Tpb: Tail position in body

Fig. 2 Subject gene linear list

2.3 Formulae of weight computation

To compare the importance of different subject strings in the same news and the importance of the same subject string in different news, we need to get the weights of subject strings. We choose frequency, position and length as the basic factors in weight computation. Among the three factors, frequency of a subject string is very important. How can we obtain the frequency of a final subject string? Because the final subject string is obtained by a series of connection and deletion, we have to keep the frequency information during connection and deletion. In other words, when connecting two shorter *SG*s to a longer *SG*, we need to keep the frequency information of the two original *SG*s. When deleting repeated *SG*s, we also need to consider the effect of deleted *SG*s on the frequency of the remained one. We use Formula 1 and Formula 2 to compute the new *SG*'s frequency, and then use Formula 3 and Formula 4 to compute the final weight of a subject string.

Formula 1. The new *SG*'s frequency after connection

The two connected *SG*s may have different frequencies and different lengths. To keep original information in new *SG*, we need to compute the new *SG*'s frequency and its length. We think that longer *SG* has bigger effect on the new *SG*'s frequency. Supposing string *C* is generated by connecting string *A* with string *B*, here *Len*(*A*), *Len*(*B*) and *Len*(*C*) stand for the lengths of the 3 strings respectively, *Freq*(*A*), *Freq*(*B*) and *Freq*(*C*) stand for their frequencies appearing in the body, we get *Len*(*C*) and *Freq*(*c*) by the following formu

$$Freq(C)=(Freq(A)*Len(A)+Freq(B)*Len(B))/(Len(A)+Len(B)) \quad (1)$$

$$Len(C)=Len(A)+Len(B)-1$$

Formula 2. The remained *SG*'s frequency after deleting repeated *SG*s

There are often several equal *SG*s before connection and we should only keep one and delete others. We select e^{1-i} as the frequency coefficient of string *i*. It is a robust coefficient after comparing with other forms. Of course, users may choose other factors for different requirements. Suppose there are *r* equal *SG*s, C_i, *i*=1,2,...,*r*. We get the frequency of the remained string according to:

© 中国科学院软件研究所 http://www.jos.org.cn

$$Freq(C)=\sum_{i=1}^{r} Freq(C_i)*e^{1-i},\ i=1,2,\ldots,r \tag{2}$$

Formula 3. The weight of subject strings

Firstly, give bigger weight for the string appearing earlier in title. This principle will help defining the most important subject about trade, which often appears earlier in news title.

Secondly, give bigger weight for longer string. Normally longer string includes more concrete information than shorter one.

Finally, give bigger weight for high-frequency string. This principle comes from such a fact that reporter often use some repeated concepts when he wants to explain his topic. And this kind of repeat may be used to decide which is more important in these concepts.

Suppose *Hpt* stands for *C*'s head character position in the title. *Freq*(*C*) and *Len*(*C*) stand for its frequency and length, just as shown in Fig.2. *Len*(*Title*) is the length of the news title. According to the three principles above, we will get the weight of subject string *C* with the following formula:

$$Weight(C)=[Len(Title)/Hpt]*Freq(C)*Len(C) \tag{3}$$

Formula 4. Normalizing the weight of subject strings

Suppose there are *q* subject strings C_j, j=1,2…*q*. C_j' weight is decided by the following formula:

$$FinalWeight(C_i)=100*Weight(C_i)/(\sum_{k=1}^{q} Weight(C_k)^2)^{1/2},\ k=1,2,\ldots,q \tag{4}$$

2.4 Processing ambiguity

In some cases, depending on the special news content, we may find that one character in title belongs to two neighbor-like subject strings. That means one string's tail is another string's head, but it is not continuous in body and does not meet the requirement of connecting them. This phenomenon prevents the completeness of subject string.

For instance, there is a piece of news as the following:

Title: 美国个人电脑在华销售将达到 60 万台

Body: 在各国纷争中国个人电脑市场的销售战中,美国在中国的销售量增长最为迅速……

In its body, the two subject strings, "美国" and "国个人电脑" can be obtained after a series of connection and deletion. Obviously, "国" belongs to "美国" and "国个人电脑". But in body, "国" as the tail of "美国" is different from "国" as the head of "国个人电脑". The ambiguity of "国" prevents the completeness of the subject string "个人电脑". So we need to take measures to correct this kind of ambiguity. Traditional method is checking and correcting the error segmentation with thesaurus, but this is just what we want to avoid. We turn to consider two important statistic factors, frequency and length. The principle is prioritizing high frequency and longer string. That is, removing the redundant characters appearing in low-frequency string (deleting this string if the string's length is less than 2 after deleting characters). If their frequencies are equal, we will delete the characters in longer *SG* because longer *SG* has more opportunity to be remained. Suppose $A=A_1J$, $B=JB_2$. *A* is in front of *B*, and the position of the first "*J*" in body is not equal to the position of the second one. We check the following condition:

$$Freq(A_1J)>Freq(JB_2) \tag{5}$$

If the condition is met, remove *J* in *B*, and decrease *B*'s length by 1. Delete *B* if *Len*(*B*<2) after removing *J*. In this example, "美国" appears twice and "国个人电脑" appears once. Then we delete "国" in the second string and get the two subject strings, "美国" and "个人电脑" ("the United States" and "personal computer").

© 中国科学院软件研究所 http://www.jos.org.cn

2.5 Main procedure

(1) Pre-Processing. Extract title and body from news. Change all of single-byte characters into double-byte ones.

(2) Establishing the character-position table. Use every title character to scan body from its beginning to end. Put the positions of each title character in the body into the data structure showed in Fig.1, and record the total times for each title character.

(3) Forming basic *SG*s. Search for all possible *SG*s from the first record to the last in Fig.1, and find their length (initially 2) and the *SG*'s head and tail character positions in the title and the body. Put all of them into the data structure showed in Fig.2.

(4) Connecting *SG*s. From bottom to top in Fig.2, compare the head character of the later *SG* with the tail character of the former *SG*. When the two characters are same and have equal position both in title and in body, delete the repeated character and connect the two *SG*s to form a new longer *SG*.

(5) Repeating step (4), connect all the possible *SG*s meeting definition 1 to form subject strings as long as possible. At the same time, compute the *SG*s frequency with formula (1) to formula (2).

(6) Computing the weight of Subject string with formula (3) and formula (4).

(7) Processing ambiguity with formula (5).

3 Experiment and Analysis

3.1 Experiment

To show the performance of this method clearly, we need to give some definitions.

Definition 2. News recall is the ratio of the news number from which subject strings can be extracted to total news number.

Definition 3. Subject recall is the ratio of the extracted "6Ws" number to all "6 Ws" number appearing in titles. The same definition can be applied to element recall, for instance, who recall, what recall and where recall.

Definition 4. New concept recall is the ratio of the number of extracted people names and organization names to the total names number appearing in titles. Currently, we only consider people names and organization names as new concepts. Of course, these new concepts are defined manually. As for other new concepts, because they are difficult to define, we do not consider them now for the fairness of experiment result.

Definition 5. Subject precision is the ratio of the number of subject strings which can express the subject of news to total extracted strings number.

Our experiment samples are 500 economic news texts mentioned in Section 1. The average length of news is 550 in Chinese characters. Our experiment is carried out in PII300, 64 RAM and Windows 98 system. Experiment method is manually checking every subject string generated by our algorithm to obtain relevant data.

To check the effect of our algorithm, we choose SE (Subject Explore) system as our contrast system. Developed by Department of Computer Science and Engineering in Shanghai Jiaotong University, SE has a subject thesaurus with 163 599 concepts, containing most of often-used concepts. The main algorithm of SE system is extracting concepts from a text with the thesaurus, based on the longest and often-used match. Then it calculates the weight of each concept according its position, frequency and length[6]. For convenience, in the following section, we name our algorithm as Algorithm A, and SE as Algorithm B. Table 2 shows part of our result.

© 中国科学院软件研究所 http://www.jos.org.cn

Table 2 Result of contrast experiment

Index	Experiment items	Algorithm A	Algorithm B
1	News recall	99.60%	100.00%
2	Mean number of subject strings per news	3.90	4 (not limited)
3	Mean length of subject strings (Chinese characters)	4.10	3.20
4	Subject recall in title	84.13%	53.67%
5	Who recall in title	97.80%	60.00%
6	What recall in title	86.20%	75.00%
7	Where recall in title	61.40%	20.00%
8	New concept recall in title	90.27%	30.00%
9	Subject precision	95.10%	75.00%
10	Subject precision of former 3 subject strings	97.50%	81.00%
11	Number of irregular strings	52	0
12	Running time (ms)	67	101

3.2 Analysis

3.2.1 General performance

Algorithm A can extract subject strings from 498 news, and its news recall reaches to 99.60%. This value is less than the 100% of Algorithm B. For Algorithm A, the mean number of subject strings per news is 3.9. For Algorithm B, this value can vary in a large scale. For convenient comparison, we choose 4. The mean length of each subject string is 4.1 (the later is 3.2). According to the theory in semantics, longer strings can express more concrete content than shorter. Therefore, the subject strings in Algorithm A are more concrete than ones in Algorithm B. Although slightly less than Algorithm B in news recall, Algorithm A can process nearly all the news with moderate number and length of subject strings.

3.2.2 Subject recall

In Algorithm A, the subject recall is 84.13%, apparently higher than 53.67% in Algorithm B. Who recall, what recall and where recall in Algorithm A are 97.80%, 86.20% and 61.40%, also higher than the corresponding data in Algorithm B.

We noticed that these data are less than the ratios of subject repeat (99.40%, 99.60%, and 95.8%) in Table 1. The main reason is that expression form in body is different from the form in title. For example, element “where”, “粤(abbreviated name of Guangdong)” in title became “广东省(Guangdong Province)”in body. To improve this performance, we may introduce area name in future study.

3.2.3 Recognizing new words

There are 290 news whose titles have people or organization names. Our algorithm can recognize 262, 90.27% of 290 news. So the algorithm is effective for most of important new concepts. Because they are difficult to be collected into thesauruses in time, the new concept recall in Algorithm B only reaches to 30%. For example, there is news titled “TOM.COM 计划首季上市”(TOM.COM Planning to Market Stocks in the First Season). In the news, “TOM.COM” is an organization name. Algorithm B is difficult to recognize it because it is a concept appearing recently. But Algorithm A is easy to identify it.

Especially, there are many abbreviated organization names in news such as “IBM”, “AOL”, and the algorithm is good at this kind of expression. But we also noticed that there are still a fraction (9.73%) of new concepts that are unrecognizable. The reason is there is some news which uses full names in the bodies instead of the abbreviated one in the titles.

© 中国科学院软件研究所 http://www.jos.org.cn

3.2.4 Subject precision

There are 95.10% of strings that are relevant to news subject in Algorithm A. For the former three subject strings (ranked in its weight), the ratio is 97.50%. Compared with Algorithm B (75% and 81%), Algorithm A has improved subject precision successfully. The reason is that all strings in Algorithm A come from title, and news title is explicit, direct and condensed to attract readers' attention almost at a glance. Our idea of extracting subject strings depending on subject repeat just exploits the nature of news.

3.2.5 Normalized degree of subject strings

Table 2 shows there are 52 irregular strings among 1952 subject strings in Algorithm A, more than 0 irregular string in Algorithm B. The reason is that the subject string in Algorithm A is obtained only by simple match, while the subject string in Algorithm B comes from regular thesaurus.

4 Conclusions

Summarizing discussions above, we can get following conclusions.

(1) Compared with other kinds of texts, Internet news often owns many new concepts. Its topic scope is unlimited, too. This is a new challenge to the traditional methods of extracting subjects based on thesaurus or linguistic rules.

(2) Exploiting the special structure of Internet news, our algorithm matches the repeated strings between news title and body. Then we can get the most important strings in the form of concepts with moderate recall and higher precision.

(3) The algorithm may be improved by additional measures to get perfect effect. For instance, to recognize more names of new organization, we can use fuzzy match to solve the problem of different expression (abbreviated names in titles and full ones in bodies).

(4) Although our research is carried out in Chinese environment, the idea can be easily applied to other languages.

(5) Our algorithm can compensate the shortage of traditional methods based on thesauruses. However, it also has some shortages such as irregular strings and little lower news recall. So combining it with traditional one is helpful and this will be one of directions in the future study.

References:

[1] CNNIC. The Statistics of Development Status about Internet in China. Statistical Report, January 2001, http://www.cnnic.net.cn (in Chinese).

[2] Gao, Jian-fang. An empirical study of CLIR at MSCN. In: Proceedings of the International Workshop ILT&CIP-2001 on Innovative Language Technology and Chinese Information Processing. German Research Center for Artificial Intelligence and Shanghai Jiao Tong University, Shanghai, 2001. 55~62.

[3] Hou, Han-qing. Research on Library Classification Thesaurus and Indexing. Beijing: Bibliography & Document Press, 1993. 281~296 (in Chinese).

[4] Hsieh, Ying-chun, Huang, Shyue-shuo. A general model of representing the content of science news using XML. In: Proceedings of the 3rd Symposium of Information Cross-Straits. Press of Taiwan Chenggong University, 2001. 143~148.

[5] Chen, Gui-lin, Wang, Yong-cheng. The research on automatic abstract of Internet information. High Technology Letters, 1999, 11(2):33~36 (in Chinese).

[6] Han, Ke-song. Research on key techniques about automatic subject distilling and indexing from texts [Ph.D. Thesis]. Shanghai Jiaotong University, 2001. 70~80 (in Chinese).

附中文参考文献:

[1] 中国互联网络信息中心.关于中国互联网络发展状况的统计.统计报告,2001.

[3] 侯汉清.当代分类法主题法索引法研究.北京:书目文献出版社,1993.281~296.

[5] 陈桂林,王永成.Internet 网络信息自动摘要的研究.高技术通讯,1999,11(2):33~36.

[6] 韩客松.中文文本主题自动提取和标引若干关键技术研究[博士学位论文].上海交通大学,2001.70~80.

© 中国科学院软件研究所 http://www.jos.org.cn

利用串匹配技术实现网上新闻的主题提取

尹中航, 王永成, 蔡 巍, 韩客松

(上海交通大学 电子信息学院,上海 200030)

摘要: 从文本中提取主题串是自然语言处理的重要基础之一.传统的提取方法主要是依据“词典加匹配”的模式.由于词典的更新速度无法同步于网上新闻中新词汇涌现的速度,而且词典的内容也无法完全涵盖网上新闻的范围,因此这种方法不适用于网上新闻的主题提取.提出并实现了一种不用词典即可提取新闻主题的新方法.该方法利用网上新闻的特殊结构,在标题和正文间寻找重复的字串.经过简单地处理,这些字串能够较好地反映新闻的主题.实验结果显示该方法能够准确、有效地提取出绝大部分网上新闻的主题,满足新闻自动处理的需要.该方法同样适用于其它亚洲语言和西方语言.

关键词:网页信息处理;网上新闻;主题提取;串匹配;词

中图法分类号: TP181 文献标识码: A

第 19 届全国数据库学术会议

征 文 通 知

由中国计算机学会数据库专业委员会主办、郑州大学和河南大学承办的第 19 届全国数据库学术会议(NDBC2002)将于 2002 年 8 月 26 日-29 日在中国河南省郑州举行.VLDB2002 将于 2002 年 8 月 20 日-23 日在香港举行,NDBC2002 将借此良机于 8 月 26 日举行 Post-VLDB,邀请国际知名专家做专题报告.

会议宗旨:本届会议将为中国大陆、香港、台湾、澳门和海外华裔数据库研究者、开发者和用户提供一个中华数据库论坛,交流有关数据库研究与应用的成果和经验,探讨数据库研究与应用所面临的关键性挑战问题和研究方向.届时国内外著名专家将到会作专题报告,主流厂商将展示他们的最新技术.会议将评选大会优秀论文和研究生优秀论文.会议正式论文将作为《计算机科学》专辑出版.我们诚征数据管理和应用领域各方面的最新成果以及有关新数据库技术、应用与方法的论文、专题讨论、演示等.

征文范围(但不限于这些领域):Web 与数据库;面向对象与对象-关系数据库系统;移动计算和数据库;WEB 缓冲技术;数据库实现技术;实时数据库系统;数据库安全性;科学与统计数据库;数据仓库和 OLAP;XML 和半结构化数据库系统;数据挖掘和知识发现;空间和时态数据库系统;文本数据库与信息检索;内容管理/知识管理;查询处理与用户界面;工作流/数据库应用;生物与基因信息系统;并行和分布式数据库系统;数字图书馆;多媒体数据库技术;数据集成和迁移;电子商务/电子政务

投稿要求:(1)论文应是未发表的研究成果,论文应包括题目、摘要、关键字、正文和参考文献.作者信息单独另纸提供,包括论文题目、作者全名、所属单位、电子邮件、通信地址、电话和传真;(2)论文中英文均可,用 Word 软件排版;论文篇幅一般不超过 6 页(A4 幅面);(3)会议论文采用网上提交方式,具体要求见会议网址: http://www.zzu.edu.cn/ndbc2002/

重要日期: 论文提交截止时间:2002 年 4 月 1 日 论文录用通知时间: 2002 年 5 月 15 日

排版稿件截止时间:2002 年 6 月 15 日

会议信息可以通过访问网站 http://www.zzu.edu.cn/ndbc2002/得到,也可以与会务组联系.

通讯地址: 450052 河南省郑州大学计算机科学系 NDBC2002 会务组

电话: 86-371-7761542(NDBC2002 会务组); 86-371-7763209(郭淑艳女士) 传真: 86-371-7761542

E-mail: ndbc2002@zzu.edu.cn

© 中国科学院软件研究所 http://www.jos.org.cn