ARTICLE IN PRESS

Future Generation Computer Systems ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

Future Generation Computer Systems

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

Power law based foundation for the measurement of discrimination information for human knowledge representation

Zheng Xu [a], Xiangfeng Luo [b,*], Yunhuai Liu [a], Lin Mei [a], Chuanping Hu [a]

[a] The Third Research Institute of the Ministry of Public Security, 201142, Shanghai, China

[b] Shanghai University, Shanghai, China

HIGHLIGHTS

- A framework on the computation of Discrimination Information.
- The proposed Discrimination Information computing algorithm reduces the computing complexity.
- The proposed method does not need any prior knowledge.
- Proposed a framework on the computation of Discrimination Information.
- The proposed Discrimination Information computing algorithm reduces the computing complexity.
- The proposed method does not need any prior knowledge.

ARTICLE INFO

Article history:
Received 27 May 2016
Received in revised form
1 September 2016
Accepted 23 October 2016
Available online xxxx

MSC:
G.3
H.1.1

Keywords:
Algorithms
Design
Theory
Discrimination information
Power law
Information theory

ABSTRACT

The discrimination information (DI) of keyword plays an important role in information retrieval and data mining. However, the measurement of DI is still a challenge because the existing methods cannot leverage the contradiction between accuracy and complexity. In this paper, a new model is proposed, does not need any prior knowledge and the computing complexity is $O(nm)$ for a collection of m documents with n keywords. Firstly, we define three types of keywords according to the document frequency spectrum, which divides the spectrum of keywords into two monotonically spectrums that can give a qualitative analysis of DI. Secondly, in order to decrease the complexity, the power law function of keywords' document frequencies is built. Thirdly, we propose an algorithm to classify keywords by using the distances between the adjacent points on the linear regression line. Finally, a piecewise function is used for computing DI according to the monotonically spectrums, which transforms DI into a scalable value to be used directly, thereby reducing the computing complexity of DI significantly. Moreover, a new weighting scheme of keywords based on DI is employed for document clustering, which shows that DI has a good prospect on the information retrieval area.

1. Introduction

It has been widely recognized that different keyword possesses diverse discrimination information (DI) in a knowledge base system. For example, "Computer" possesses a lower DI than "CPU" in the computer field. "Example Learning" possesses a higher DI than "Intelligence" in the area of artificial intelligence. In reality, DI has a wide range of applications including semantic annotations for Web pages [1–3], discovery of semantic community [4–6], documents clustering/classification [7–9], e-learning technology [10,11,4], etc. In addition, DI is important for web search [12–15], which can be used for query expansion to help users find more relevant information. Therefore, how to compute DI is a basic problem for information retrieval and data mining.

In [16], Salton et al. regarded DI as a measurement of the variation in the average similarity between documents in a collection. A good discriminator is an assigned keyword which can reduce the average similarity between documents. In contrast, a poor discriminator increases the inter-document similarity. Unfortunately, the computing complexity of DI is proportional to $O(nm^2)$ for a collection of m documents with n keywords, which is unpractical to be used directly for a collection containing large documents. Cai [17] uses information theory to compute DI. In that work, the discrimination information of a keyword refers to the amount of infor-

* Corresponding author.

http://dx.doi.org/10.1016/j.future.2016.10.021
0167-739X/

mation conveyed by a keyword in support of a certain category of documents and rejecting other categories. An informative keyword should have a high capability of categorizing document.

In this paper, DI refers to the capability of semantic discrimination conveyed by a keyword in support of an expected information need (e.g., Salton et al. focus on document clustering and Cai focuses on document classification). To understand the meaning of the proposed definition more clearly and precisely, we provide two examples below.

The first example is about document clustering. Document clustering is to automatically group text documents into clusters so that documents within a cluster have high similarity with each other, but are dissimilar to documents in another cluster. In this example, the expected information need can be summarized as obtaining "accurate document clustering results". According to this information need, keywords with high DIs should contribute more to the similarity of the documents within a cluster than to other clusters. Since each document is often represented as a vector in the existing document clustering algorithms, the document frequency of a keyword may be useful for the computation of DI. A keyword with a high document frequency may have low DI since it may augment the similarity of documents within different clusters. Similarly, a keyword with low document frequency also has low DI since it may reduce the similarity of documents within a cluster.

The second example is about query suggestion. In query suggestion, a set of concepts related to the query is suggested to help the user find what she/he really needs, thereby improving user's search experience and retrieval effectiveness. The expected information need can be summarized as obtaining "accurate search results" in this case. Keywords with high DIs should provide sufficient user information needs for effectively retrieving relevant pages. Similar to example one, we find the page counts of keywords may affect DI. Keywords with high page counts may not be useful for searching, as high page counts will bring up a large number of search results, which may reduce the precision of search results. On the other hand, keywords with low page counts may also have low DIs since they may reduce the recall of search results due to their rare occurrence.

As seen from the above examples, the computation of DI is not a simple issue. It is still a challenge for the existing methods, such as the high computing complexity of Salton et al.'s discrimination value model [16]. In this paper, we do not aim at giving a common computation model of DI since it is unpractical. Instead, we concentrate on the following situation: suppose a document space D in which each document d is represented as a n-dimensional vector, $d = (w_1, w_2, \ldots, w_n)$, where w_n represents the weight of the nth keyword. We want to compute DI in support of an expected information need on the document space D. In particular, the following two important questions arise in this context: (1) how to find appropriate factors to influence DI; (2) how to construct a function integrating these factors to compute DI. To address these two issues, in this paper, we propose a method using the power law function of keywords' document frequencies. Our method consists of the following four major steps. First, three types of keywords are defined according to their document frequency spectrum, which divides the spectrum of keywords into two monotonically spectrums that can facilitate a qualitative analysis of DI. Second, in order to decrease the computing complexity, the power law function of keywords' document frequencies is built. Third, an algorithm is proposed to classify keywords into three types, by using the distances between the adjacent keywords on the linear regression line of the power law function. Finally, entropy is used to construct a piecewise function for computing DI according to the monotonically spectrums, which transforms DI into a scalable value to be used directly, thereby reducing the computing complexity significantly. The major contributions of our work are summarized as follows.

(1) A framework on computing DI is proposed including the classification of keywords based on their document frequency spectrum, and the power law feature of keyword frequencies.
(2) An algorithm of computing DI is proposed, which transforms the problem to the study of distances between the adjacent points in the linear regression line of keywords. The proposed DI computing algorithm reduces the computing complexity from $O(nm^2)$ by Salton et al.'s model to $O(nm)$ of our model for a collection of m documents with n keywords. Experimental results show that Salton et al.'s model can be replaced by ours with a very low error rate.
(3) The proposed method does not need any prior knowledge such as the distribution of keywords. The document frequency spectrum which is easily obtained from document collections is used to compute DI. Entropy is used to construct a piecewise function for computing DI, which transforms DI into a scalable value that can be obtained directly.
(4) Extensive experiments are conducted to evaluate the proposed method. Four data sets are used to evaluate the accuracy of the qualitative analysis on DI. The result of Salton et al.'s model is used as the benchmark since its high computing complexity ensures the accuracy of computing DI. The high correlation coefficient between our model and Salton et al.'s model shows that our model can effectively compute DI in a highly accurate manner. Moreover, in order to demonstrate the utility of our work, a new weighting scheme of keywords based on DI is employed for document clustering. Experimental results confirm that our model outperforms the term frequency-inverse document frequency on the results of document clustering by a wide margin.

The rest of the paper is organized as follows. In the next section, the related work is given. Section 3 introduces the power law function of keywords. Section 4 discusses the relation between DI and the power law function of keywords. The algorithms for identifying the general keywords and the minimum rank keywords are proposed in Sections 5 and 6, respectively. The function of computing DI is proposed in Section 7. An application of using DI on document clustering is introduced in Section 8. The last section gives the conclusions of our work.

2. Related work

Formalization and quantification of the intuitive notion of discrimination measures have long been a major challenge for computing science, and an intriguing problem for other domains. Discrimination measures may be first proposed by [16]. Salton et al. [16] believed that DI is a measurement of the variation in the average similarity between documents in a collection. A good discriminator is an assigned keyword which will reduce the average similarity between documents. Salton et al. [16] selected 450 documents from the area of medicine as the experimental data. According to the document frequencies of words, totally 4726 keywords in 450 documents have been divided into different classes. For each class of keywords, the average rank of the corresponding keywords is given according to the value of DI in descending order (i.e., the lower, the better). The experimental results show that keywords with high/medium/low document frequencies possess lowest/high/low DI because they are the worst/best/poor semantic discriminators. The keywords with very low or very high document frequency possess rather poor average ranks, and the best semantic discriminators are those keywords whose document frequency is neither too low nor too high. Some other methods use document frequency (df) to compute DI because it can be easily obtained from document collections. These methods include inverse document frequency (idf) [18], mutual information [19], information gain [20], relevancy score [21], and

Please cite this article in press as: Z. Xu, et al., Power law based foundation for the measurement of discrimination information for human knowledge representation, Future Generation Computer Systems (2016), http://dx.doi.org/10.1016/j.future.2016.10.021

Table 1
The details of data sets used to measure the DI of keywords.

The category of reuters	The number of news
Environment news (1)	977
Environment news (2)	2200
Health news	2000
Internet news	2000

so on. Though the computing complexity of using document frequency (df) to compute DI is simple, the accuracy of these methods is still low.

Different from the above methods, some methods use information theory for discrimination measures. Topsoe [22] presented two measures of discrimination between two probability measures P and Q named capacitor discrimination and triangular discrimination. These two measures use the functions from divergence measures. Similar work which uses information divergence for discrimination measures include [23,24], which consider the discrimination measures as an information theory problem. Recently, Cai [17] regarded DI as a measurement of the divergence from information theory. Cai [17] uses information theory to compute DI. The basic idea proposed by Cai is that the extent of the contribution that a term makes may hence be used as a device for measuring the informativeness of that term. The underlying mathematical structures that enable the computation are divergence measures drawn from information theory.

Based on the analysis of the above methods, we can see that two factors are important in the computation of DI, including:

(1) Low computing complexity. For DI to be practically usable on a large scale document space such as the Web search and document recommendation, the computing complexity should be low enough for these applications.
(2) Independence on the priori knowledge. The priori knowledge such as distribution of keywords on different classes, the hierarchical structure of keywords, can be hardly obtained by the ordinary text mining tasks. Thus, the methods for computing DI should rely on such priori knowledge as little as possible.

3. The power law function of keywords

In this section, we first discuss the power law function of keywords. We next introduce the linear regression which leads us to identify whether a function follows a power law or not.

3.1. The data sets

As mentioned in Section 2, when using Salton et al.'s model, the number of the documents and the keywords should be small enough because the computing complexity of Salton et al.'s model is $O(nm^2)$. Table 1 lists some categories of news downloaded from www.reuters.com as the data sets.

3.2. Linear regression of the power law function

According to the definition in [25,26], a power law is a function with the form $f(x) = \alpha x^{-\beta}$, where α and β are constants. In order to measure the frequency of variable X, a probability density function (PDF) can be used, that is, $P(X = x)$. In addition, a complementary cumulative distribution function (CCDF) (that is, $P(X > x)$) can also be used to measure the frequency of variable X.

Moreover, the issue of investigating whether or not PDF follows a power law relies on a commonly used method, namely, linear regression [27]. The accuracy of the approximation is indicated by the coefficient of the determination R^2. Following [25], whenever we say that a function follows a power law, we mean that $R^2 \geq 0.9$. Before we test the power law feature of keywords, one basic definition is given.

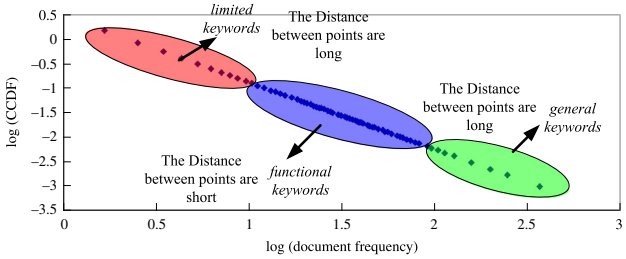

Fig. 1. Three parts corresponding to the general keywords, functional keywords, and limited keywords. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Definition 1. A set of document frequencies ($df(D)$) is the document frequencies of all keywords in document collection D by ascending order, that is, $df(D) = \{df_1, df_2, \ldots, df_{|df(D)|}\}$. Especially, $df_1 < df_2 < \cdots < df_{|df(D)|}$.

In this section, the linear regression of CCDF on the document frequency of keywords is measured on the data sets for checking if the distribution of the document frequency follows power law or not. Table 2 shows the linear regression results for all the data sets. It is unsurprising that the document frequencies of keywords on all the four datasets follow a power-law function.

In fact, the power law features of keywords' document frequencies have been investigated for a long history. Zipf's law [28], an empirical law formulated using mathematical statistics, refers to the fact that many types of data studied in the physical and social sciences can be approximated with a Zipfian distribution, one in the family of related discrete power law probability distributions. In the next section, we want to analyze whether this feature can be used to compute DI or not.

4. DI and power law function: theoretic analysis

In this section, a theoretic analysis on the relation between DI and the power law function of keywords' document frequencies is conducted. Finally, the method for computing DI is proposed.

4.1. On the distance feature of linear regression line

Since we use the document frequencies of keywords as the object of linear regression line, a definition is given below:

Definition 2. $Point_i$ is the ith point on the linear regression line, which is a group of keywords with the same document frequency, that is, $Point_i = \{x | \forall x \rightarrow df(x) = df_i\}$, where x is a keyword of document collection D. Herein, $df(Point_i) = df_i$.

From Definition 2, $Point_i$ means the ith group of keywords with the document frequency df_i. After giving the definition of $Point_i$, we give the definition of distance between the adjacent points.

Definition 3. $Dis(Point_i, Point_j)$ is the distance between $Point_i$ and $Point_j$ on the linear regression line.

According to Definition 3, we can infer that the average distance between $point_i$ and $point_j$($Ad(point_i, point_j)$) is equal to $\sum_i^{j-1} Dis(Point_k, Point_{k+1})/(j-i)$.

If all of the points in the linear regression line are vertically projected to the fitting line of the linear regression (shown in Fig. 1), an important characteristic can be obtained as follow.

Please cite this article in press as: Z. Xu, et al., Power law based foundation for the measurement of discrimination information for human knowledge representation, Future Generation Computer Systems (2016), http://dx.doi.org/10.1016/j.future.2016.10.021

Table 2
The linear regression results of data sets listed in Table 1.

Category of reuters	Coefficient of determination, R_2	Follow a power law or not
Environment news (1)	0.9591	Yes
Environment news (2)	0.9497	Yes
Health news	0.9786	Yes
Internet news	0.9730	Yes

Characteristic 1. The average distance in the both ends of the linear regression line is higher than that in the middle of the linear regression line. The average distance feature of keywords' document frequencies is similar to the document frequency spectrum of DI.

In Fig. 1, the distances between the adjacent points in the green and red eclipses are longer than that in the blue eclipse. In other words, the average distance in the blue eclipse is higher than that in the green and red eclipses. Especially, the document frequencies of the keywords in the blue eclipse are medium. Salton et al. [16] regarded that the keywords with high/medium/low document frequencies possess lowest/high/lower DI because they are the worst/best/poor semantic discriminators. This result is similar to the average distance feature of the linear regression line. The points (words) in the blue eclipse possess higher DI than those in the red and the green eclipses. Characteristic 1 may be useful for computing DI since the different average distance feature in different eclipses. We give below some theoretic analysis on the average distance feature of linear regression line. We are interested in knowing which reason causes it and which parameters impact on it. According to Characteristic 1, the distance feature is similar to that of DI. Thus, the distance feature can be used to compute DI. Characteristic 2 shows the factors impacting on the distance between the adjacent points on the linear regression line.

Characteristic 2. $Dis(Point_i, Point_{i+1})$ is determined by PDF_{i+1} and $df_{i+1} - df_i$, that is, $Dis(Point_i, Point_{i+1}) = f(PDF_{i+1}, df_{i+1} - df_i)$.

Proof. Suppose $Point_i$ with df_i, PDF_i, and $CCDF_i$, the linear regression line is $y = Ax + B$, where x is df_i and y is $CCDF_i$; and the line vertical to the linear regression line is $y = -x/A + C$.

Put $(df_i, CCDF_i)$ to $y = -x/A + C$, then $C = CCDF_i + \frac{df_i}{A}$, and the vertical line is

$$y = -\frac{x}{A} + CCDF_i + \frac{df_i}{A}. \tag{1}$$

The vertical projection point of $Point_i$ is computed by

$$\begin{cases} y = Ax + B \\ y = -\dfrac{x}{A} + CCDF_i + \dfrac{df_i}{A}. \end{cases} \tag{2}$$

The result of Eq. (2) is

$$\begin{cases} x = \dfrac{A * CCDF_i + df_i - AB}{A^2 + 1} \\ y = \dfrac{A^2 * CCDF_i + A * df_i + B}{A^2 + 1}. \end{cases} \tag{3}$$

Based on the coordinate of $Point_i$ and $Point_{i+1}$, the distance $Dis(Point_i, Point_{i+1})$ between them is given by Eq. (4) (see Box I). Since $CCDF_{i+1} - CCDF_i$ is equal to PDF_{i+1}, Eq. (4) can be replaced by

$$\sqrt{\frac{\left(A * PDF_{i+1} + (df_{i+1} - df_i)\right)^2 + \left(A^2 * PDF_{i+1} + A(df_{i+1} - df_i)\right)^2}{\left(A^2 + 1\right)^2}}. \tag{5}$$

Thus, we can see that the distances between the adjacent points in the linear regression line is determined by PDF_{i+1} and $df_{i+1} - df_i$ via Eq. (5).

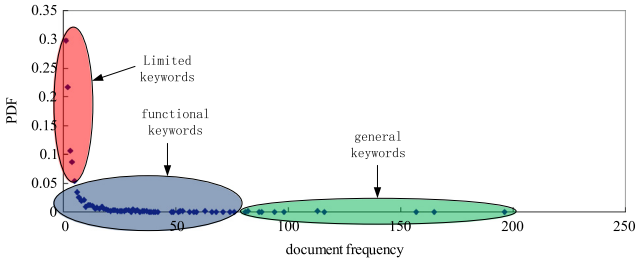

Fig. 2. The PDF result of "environment new (1)" data set.

4.2. Classification of keywords

According to Characteristic 1, we denote the keywords in the document space as general keywords, functional keywords, and limited keywords, respectively.

Deduction 1. According to Characteristic 1, among all of the keywords, the medium discriminator keyword (mdk) is the one that is the boundary between the best and the worst discriminators in Salton et al.'s model. For example, in Fig. 1, mdk is the point as the boundary between the green and blue eclipse.

Deduction 2. According to Characteristic 1, among all of the keywords, the secondary discriminator keyword (sdk) is the one that is the boundary between the poor and the worst discriminators in Salton et al.'s model. For example, in Fig. 1, sdk is the point as the boundary between the red and blue eclipse. According to Deductions 1 and 2, we can obtain characteristic 3 below.

Characteristic 3. The document frequency of medium discriminator keyword is higher than that of secondary discriminator keyword, which means $|D| \geq df(mdk) \geq df(sdk) \geq 0$, where $|D|$ means the number of documents in document space D.

This characteristic is easy to be understood, which is caused by the medium/low/high document frequency of best/poor/ worst discriminators.

Definition 4. Among all of the keywords, the general keywords (GK) are those whose document frequencies are higher than medium discriminator keyword, which means $\forall k_i \in GK \rightarrow df(k_i) \geq df(mdk)$. The general keywords are the worst discriminators of all keywords, which is with the high document frequencies. For example, the points in the green eclipse of Fig. 1 belong to the general keywords.

Definition 5. Among all of the keywords, the limited keywords (LK) are those whose document frequencies are lower than secondary discriminator keyword, which means $\forall k_i \in LK \rightarrow df(k_i) \leq df(sdk)$. The limited keywords are the poor discriminators of all keywords, which is with the low document frequencies. For example, the points in the red eclipse of Fig. 1 belong to the limited keywords.

Please cite this article in press as: Z. Xu, et al., Power law based foundation for the measurement of discrimination information for human knowledge representation, Future Generation Computer Systems (2016), http://dx.doi.org/10.1016/j.future.2016.10.021

$$\sqrt{\frac{\left(A(CCDF_{i+1}-CCDF_i)+(df_{i+1}-df_i)\right)^2+\left(A^2(CCDF_{i+1}-CCDF_i)+A(df_{i+1}-df_i)\right)^2}{\left(A^2+1\right)^2}}. \quad (4)$$

Box I.

Definition 6. Among all of the keywords, the functional keywords (FK) are those whose document frequencies are higher than medium discriminator keyword and lower than secondary discriminator keyword, which means $\forall k_i \in FK \rightarrow df(mdk) \geq df(k_i) \geq df(sdk)$. The functional keywords are the best discriminators of all keywords, which is with the medium document frequencies. For example, the points in the blue eclipse of Fig. 3 belong to the functional keywords.

According to Definitions 4, 5, and 6, two deductions about distance feature can be given as follows.

Deduction 3. The average distance of general keywords is higher than that of functional keywords, that is, $Ad(mdk, Point_n) > Ad(sdk, mdk)$, where the document frequency of $Point_n$ is the highest document frequency of $df(D)$.

Deduction 4. The average distance of limited keywords is higher than that of functional keywords, that is, $Ad(Point_1, sdk) > dd(sdk, mdk)$, where the document frequency of $Point_1$ is the lowest document frequency of $df(D)$.

4.3. On the distance feature of general keywords and limited keywords

In this section, we want to analyze the reason of the higher average distance for general keywords and limited keywords than that of functional keywords. The result of *PDF* is given in Fig. 2. From this figure, we can see that $df_{i+1}-df_i$ of the general keywords are higher than those of the functional keywords and the limited keywords.

Deduction 5. The higher $df_{i+1}-df_i$ of general keywords than that of functional keywords causes its higher average distance in the linear regression line than that of functional keywords, that is, $\frac{\sum_l^{|D|-1}(df_{i+1}-df_i)}{|D|-l} > \frac{\sum_k^{l-1}(df_{j+1}-df_j)}{l-k}$, where $df(Point_l) = df(mdk)$ and $df(Point_k) = df(sdk)$.

In order to verify Deduction 5, we compute $df_{i+1}-df_i$ for the general/functional/limited keywords on the four data sets. As shown in Table 3, we can see that on average $df_{i+1}-df_i$ in the general keywords is indeed higher than those of the functional keywords and the limited keywords.

From Fig. 2, we can also see that the PDF_i of the limited keywords are higher than those of the functional keywords.

Deduction 6. The higher PDF_i of limited keywords than that of functional keywords causes its higher average distance in the linear regression line than that of functional keywords, that is, $\frac{\sum_1^{k-1}(df_{i+1}-df_i)}{k-1} > \frac{\sum_k^{l-1}(df_{j+1}-df_j)}{l-k}$, where $df(Point_l) = df(mdk)$ and $df(Point_k) = df(sdk)$.

In order to verify Deduction 6, we have computed the PDF_i on the four data sets. The results are shown in Table 4. From this table, we can see that the PDF_i of the limited keywords are indeed higher than those of the functional keywords and the general keywords.

4.4. Computing DI: the proposed method

In the above three sections, we analyze the average distance feature of different groups of keywords. But the classification of keywords (general keywords, functional keywords, limited keywords) is not enough for computing DI. Fig. 3(a) gives the illustration of the classification of keywords against the document frequency spectrum.

Characteristic 4. The keywords on the right/left of the document frequency spectrum are approximate to a monotonically increasing/decreasing function with document frequency.

Since the general keywords are with high document frequencies, the keywords in the blue eclipse in Fig. 3 can be moved to the left on the document frequency spectrum. This step is illustrated by Fig. 3(b). Thus, the document frequencies of the general keywords are transformed to be lower than those of the limited keywords.

Deduction 7. The minimum rank keyword (mrk) is the one whose rank is the highest of all the keywords. In other words, the DI of mrk is the highest of all the keywords.

After finding the highest value of DI, we can construct a piecewise function based on the minimum rank keyword of the document frequency spectrum. A monotonically decreasing/increasing function with document frequency can be obtained by the keywords on the right/left of the spectrum of keywords, as illustrated by Fig. 3(c). Consequently, our proposed DI computation has the following three steps.

(1) **Identifying the general keywords** (computing $df(mdk)$). This step is to identify the general keywords. That is, the lowest document frequency of the general keywords should be identified correctly. We shall thus get the document frequency of medium discriminator keyword (mdk).
(2) **Identifying the minimum rank keyword** (computing $df(mrk)$). Since the function for computing DI is a piecewise function based on the minimum rank point of document frequency spectrum, it is necessary to identify the keyword with the minimum DI. We shall thus get the document frequency of minimum rank keyword (mrk).
(3) **Building a piecewise function to compute DI**. With the above two steps, this step constructs a piecewise function based on the minimum rank point of document frequency spectrum.

In essence, our proposed method aims to turn the problem of computing DI into the study of the distances between the points in the linear regression line of keywords. Clearly, it is much cheaper to compute the distance between the adjacent points than employing Salton et al.'s model. In this way, Salton et al.'s complicated model can be replaced by a relatively simple model, with a low error rate.

5. Identifying the general keywords

5.1. Algorithm

According to Characteristic 1, the average distance feature of the three types of keywords inspires us to use it for identifying the general keywords. In this section, an algorithm is put forward to identify the general keywords based on the linear regression line of keywords. The steps of our algorithm are as follows.

Please cite this article in press as: Z. Xu, et al., Power law based foundation for the measurement of discrimination information for human knowledge representation, Future Generation Computer Systems (2016), http://dx.doi.org/10.1016/j.future.2016.10.021

Table 3
The average $df_{i+1} - df_i$ of the general keywords, the functional keywords, and the limited keywords in the four data sets.

The category of reuters	General keywords	Functional keywords	Limited keywords
Environment news (1)	9.14	1.37	1
Environment news (2)	11.93	1.49	1
Health news	26.52	1.4	1
Internet news	21.68	1.73	1

Table 4
The PDF_i of the general keywords, the functional keywords, and the limited keywords in the four data sets.

The category of reuters	General keywords	Functional keywords	Limited keywords
Environment news (1)	0.000718	0.002612	0.080194
Environment news (2)	0.000581	0.002434	0.062565
Health news	0.000602	0.002672	0.073808
Internet news	0.000657	0.002448	0.056526

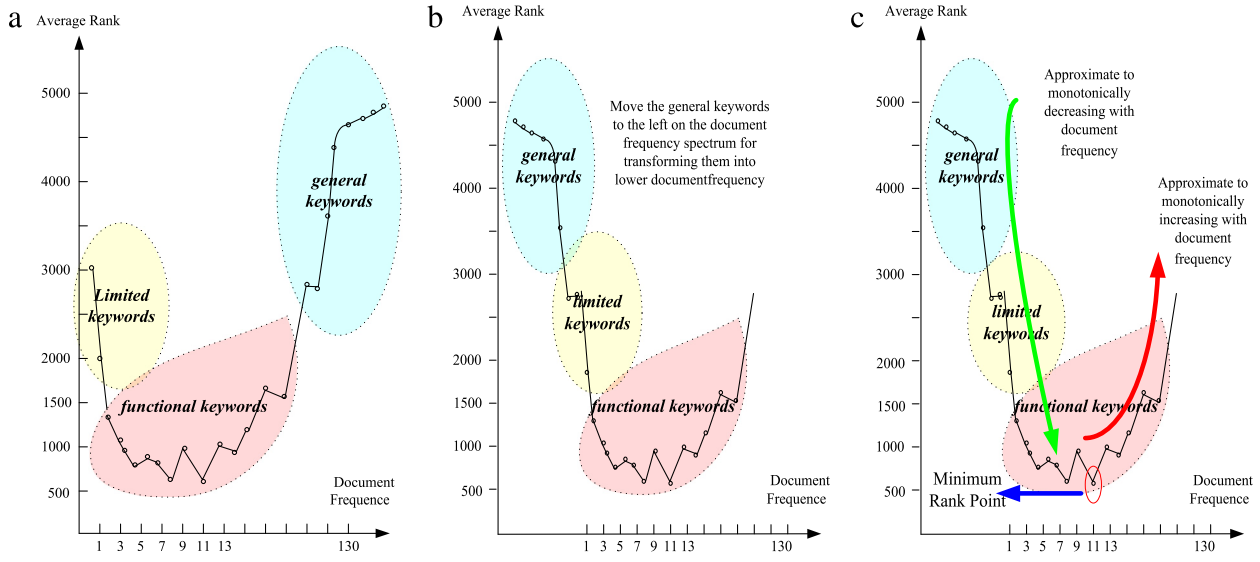

Fig. 3. The illustration of the proposed DI model. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

(1) Computing the distance between adjacent points $Dis(Point_i, Point_{i+1})$ in the linear regression line according to Eq. (5);
(2) Computing the average distance of linear regression line $Ad(Point_1, Point_n)$ according to Deduction 1:

$$Ad(Point_1, Point_n) = \sum_{1}^{n} dis(Point_i, Point_{i+1}) \Big/ n - 1. \quad (6)$$

(3) Obtaining a binary digit (bd) of each $Point_1$ via:

$$bd(Point_i) = \begin{cases} 1, & \text{if Dis}(Point_i, Point_{i+1}) \le Ad(Point_1, Point_n) \\ 0, & \text{else.} \end{cases} \quad (7)$$

(4) Obtaining a binary string by computing $bd(Point_i)$ of each $Point_i$.

According to the average distance feature of Characteristics 1, a special feature of the binary string is given below:

Characteristic 5. The binary string feature of $Ad(Point_1, Point_n)$ is similar to that of DI. There are so many '0's in the two ends of the binary string and so many '1's in the middle, the binary string can also be divided into three parts like keywords.

Thus, the binary string is seen as an operational objective to identify the general keywords, which can be obtained by the follow steps:

(1) Initialize a sliding-window. Set its size as the length of the longest '0' substring of the binary string;
(2) Compute the proportion between '1' and '0' in the binary string, which we denote as R0;
(3) Put the sliding-window at the beginning of the binary string and move the sliding-window from the left to right until the proportion between '1' and '0' in the sliding-window is greater than R0;
(4) Set the string in the left side of the sliding window as the general keywords.

An example of the above procedure is given below. According to Characteristic 5, suppose the binary string is

00001000011111111111101111111111111110000000.

Step (1): Set the size of sliding-window as 7;
Step (2): R0 = 26:16;
Step (3): Move the sliding-window (marked as red character) from left to right, until the proportion between '1' and '0' in the sliding-window is greater than R0:

0000100001111111111110111111111111110000000
0000100001111111111110111111111111110000000
0000100001111111111110111111111111110000000
0000100001111111111110111111111111110000000

Step (4): Mark the general keywords as red color.

0000100001111111111110111111111111110000000

Table 5
The error rate of the algorithm for identifying the general keywords on each data set.

Category of reuters	Error rate
Environment news (1)	0.32%
Environment news (2)	0.34%
Health news	0.46%
Internet news	0.12%

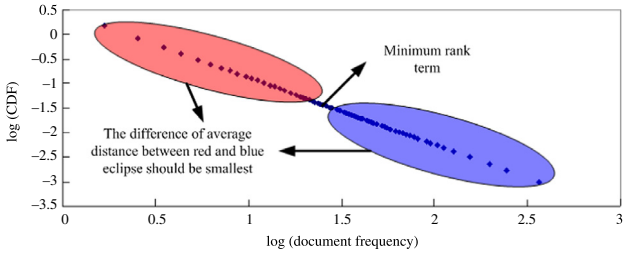

Fig. 4. An illustration for the main idea of the algorithm for identifying the minimum rank keyword. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Step (5): Mark the limited keywords as blue color.

0000100001111111111101111111111110000000

Through our algorithm, we can convert the problem of identifying the general keywords from Salton et al.'s model to a simpler process of computing the distance between the adjacent points of keywords. Moreover, our algorithm does not need any prior knowledge.

5.2. Evaluating the algorithm

In order to evaluate the algorithm of identifying the general keywords, its complexity should be analyzed first. The proposed algorithm reduces the complexity from Salton et al.'s $O(nm^2)$ to $O(nm)$. The proposed method only compute the document frequency of each keyword, which is $O(nm)$.

While the complexity of our proposed algorithm is lower than Salton et al.'s model, we also need to check the error rate between the keywords classified by our proposed algorithm and that of Salton et al.'s model. Suppose the set of general keywords identified by our method is $GK_1 = \{k_1, k_2, \ldots, k_{|GK_1|}\}$, and that by Salton et al.'s model is $GK_2 = \{k_1, k_2, \ldots, k_{|GK_2|}\}$. The error rate (er) is computed by

$$er = |GK_1 - GK_2| / m \tag{8}$$

where m denotes the number of the keywords in the document collection D. $|GK_1 - GK_2|$ means the number of the elements in the difference set between GK_1 and GK_2.

Table 5 lists the error rate of our proposed algorithm on the four data sets. The error rates of four data sets are all lower than 0.5%, which shows that the error rate is generally low enough for identifying the general keywords correctly.

6. Identifying the minimum rank keyword

It is necessary to identify the minimum rank keyword because the function for computing DI is a piecewise function based on it. Similar to the algorithm for identifying the general keywords, the distance between the adjacent points in the linear regression line is also used due to its lower complexity than that of Salton et al.'s model. Moreover, we provide a test to check if the algorithm for identifying the minimum rank keyword is correct or not.

Table 6
The error rate of the algorithm for identifying the minimum rank keyword on each data set.

Category of reuters	Error rate
Environment news (1)	0.25%
Environment news (2)	1.5%
Health news	1.6%
Internet news	1.9%

6.1. Algorithm

The main idea of the algorithm lies in that the difference of the average distance between the left and the right points of the minimum rank point should be as small as possible. As illustrated by Fig. 4, the keywords on the left and right side of the minimum rank point are in the red and blue eclipse, respectively. The difference of the average distance between the red and blue eclipse should be the smallest. The main steps for identifying the minimum rank keyword are as follows.

(1) Compute the difference of average distance (Da) between the left and right points of $Point_i$ by

$$Da(Point_i) = \left| \frac{Dis(Point_1, Point_i)}{i-1} - \frac{Dis(Point_i, Point_n)}{n-i} \right|. \tag{9}$$

(2) Compute $Da(Point_i)$ of each $Point_i$;
(3) Sort $Point_i$ according to their $Da(Point_i)$;
(4) Select $Point_i$ with the lowest $Da(Point_i)$ as the minimum rank keyword.

6.2. Evaluating the complexity and error rate

The complexity of the proposed algorithm for identifying the minimum rank keyword is $O(l^* \log l)$, as opposed to $O(nm^2)$ by Salton et al.'s model. While the complexity of the proposed algorithm is lower than that of Salton et al.'s model, we also need to check on the error rate. Similar to Eq. (8), suppose the minimum rank keyword identify by our method is mrk_1, and that by Salton et al.'s method is mrk_2. We compute the proportion of keywords from $df(mrk_1)$ to $df(mrk_2)$. Table 6 lists the error rate of our algorithm on the four data sets. The highest error rate of four data sets is 1.9%, which shows that the error rate is generally low enough for identifying the minimum rank keyword correctly.

7. Computing DI

As mentioned in Section 4, the function for computing DI is a piecewise function based on the minimum rank point of document frequency spectrum. A monotonically decreasing/increasing function with the document frequency of keywords can be gained according to the left/right side of the spectrum shown in Fig. 3. The document frequency of keywords is an important parameter to compute DI. In addition, the function should be a piecewise one based on the minimum rank keyword. As DI can be obtained by Salton et al.'s model with a low number of documents and keywords, we shall use the results of Salton et al.'s model as a benchmark to evaluate the accuracy of the constructed function.

7.1. DI of keywords

In information theory, entropy is a measure of the uncertainty associated with a random variable [26]. Inspired by entropy, DI of k_i can be computed by

$$DI(k_i) = (df(k_i)/|D|) * \log(|D|/df(k_i)) \tag{10}$$

where DI of k_i is a piecewise function based on $df(k_i) = {}^{|D|}/_{e}$, which means $DI(k_i)$ is a monotonically increasing function when $df(k_i) = {}^{|D|}/_{e}$; e is a constant.

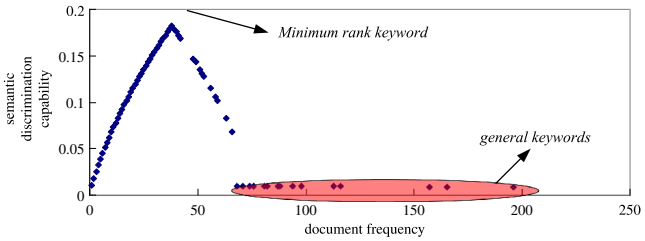

Fig. 5. The experiments on document frequency and DI of keywords.

7.2. Revising the document frequency

Since DI is a piecewise function based on the minimum rank keyword in a document frequency spectrum, we need to go through two steps before using Eq. (10) to compute DI directly.

(1) Assign the general keywords with a lower document frequency than those of the limited keywords. Since the general keywords have the lowest DI, we should give them the lowest document frequency.
(2) Assign a low document frequency to the keywords on the right side of the minimum rank keyword since DI is a piecewise function based on the minimum rank keyword in a document frequency spectrum, the keywords on the right side of the minimum rank keyword should be given a low document frequency.

The function for revising the document frequency of the general keywords is as follows:

$$Revisingdf(k_i) = \frac{|D| - df(k_i)}{|D|}, \quad k_i \in GK \tag{11}$$

where $|D|$ denotes the number of documents in document space D. Through Eq. (11), the revised document frequencies of general keywords can be obtained. Thus Eq. (10) can be used to compute DI of the general keywords directly. The function for revising the document frequency of the keywords on the right side of the minimum rank point is as follows.

$$Revisingdf(k_i) = \begin{cases} 2 * df(mrk) - df(k_i), & \text{if } (2 * df(mrk) \\ -df(k_i)) \in \left[0, \dfrac{df(k_i)}{|D|}\right] & \\ 1, & \text{else.} \end{cases} \tag{12}$$

Through Eq. (12), the revised document frequencies of the keywords on the right side of the minimum rank keyword can be obtained. Thus Eq. (10) can be used to compute DI of these keywords directly.

7.3. Computing DI

According to Eq. (10), the function for computing DI is as follows:

$$DI(k_i) = \left(\frac{Revisingdf(k_i)}{|D|}\right) * \log 2 \left(\frac{|D|}{Revisingdf(k_i)}\right). \tag{13}$$

In order to judge whether the formula is correct or not, it is necessary to test whether the actual keywords with high document frequency are indeed of low DIs. Fig. 5 shows the experimental result of DI using Eq. (13) on the "environment news (1)" data set. The horizontal axis refers to the document frequency of the keywords and the vertical axis refers to DI. It can be seen from Fig. 5 that DI of the general keywords is the lowest and DI of other keywords are piecewise according to the minimum rank keyword.

7.4. Evaluating the function for computing DI

In statistics, the Pearson correlation coefficient is a common measure of the correlation between two variables X and Y [29].

Table 7
The correlation coefficients of the data sets.

Category of reuters	Correlation coefficient
Environment news (1)	0.9975
Environment news (2)	0.9995
Health news	0.9999
Internet news	0.9996

The formula is given below:

$$\rho = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{X_i - \mu_X}{\sigma_X}\right)\left(\frac{Y_i - \mu_Y}{\sigma_Y}\right) \tag{14}$$

where $\frac{X_i - \mu_X}{\sigma_X}$, μ_X and σ_X are the standard score, population mean, and population standard deviation, respectively. According to Eq. (14), the evaluation standard for computing DI can be put forward as follows.

(1) Rank keywords according to their DI computed by Salton et al.'s model;
(2) Compute the average rank of keywords with the same document frequency, the result of which is denoted as Rank(x);
(3) Rank keywords according to their DI computed by Eq. (13);
(4) Compute the average rank of keywords with the same document frequency, the result of which is denoted as Rank (y);
(5) Multiply Rank(x) and Rank (y) by their proportions;
(6) Compute the correlation coefficient between Rank (x) and Rank (y).

Table 7 lists the results on the four data sets. The correlation coefficient of each data set is high, which means that the function to compute DI is quite accurate. We now compare our model with Salton et al. [16] and Cai [17] in terms of similarities and differences.

(1) **The proposed model vs. Salton et al.'s model**
As for the similarity, both models use document frequency which can be easily obtained from the documents collection. In addition, both methods focus on the problem of computing the discrimination information of the keywords/terms. Finally, the results of the two methods are similar, as the correlation coefficient between the two methods on each data set is rather high, as shown in Table 7. As for the difference, Salton et al.'s method uses the variation in the average pairwise document similarity as a computation to DI, with an expensive $O(nm^2)$ computing complexity for m documents with n keywords. Usually, m is large, which makes Salton et al.'s model unpractical to be used directly for a large collection of documents. In contrast, our proposed model reduces the computing complexity to $O(nm)$. Meanwhile, our model uses the document frequency spectrum and the power law function of keywords to compute DI, which transforms the DI into a scalable value, hence can be used directly. Note that the technique of computing DI can be applied to data mining tasks such as document clustering. In the next section we also show that DI has a good prospect in the information retrieval area.
(2) **The proposed model vs. Cai's model**
As for the similarity, both methods focus on the problem of computing the semantic discrimination capabilities of the keywords/terms. Both models adapt functions from information theory to compute DI. Cai formally interprets the discrimination information conveyed by a term and points out some problems in applying the measures in practice. As for the difference, Cai's model uses some divergence function for computing DI. In contrast, our proposed method use the power law feature of document frequency, and can be used for text mining related tasks easily.

Please cite this article in press as: Z. Xu, et al., Power law based foundation for the measurement of discrimination information for human knowledge representation, Future Generation Computer Systems (2016), http://dx.doi.org/10.1016/j.future.2016.10.021

Table 8
Summary description of document sets.

Data set	Source	Number of documents	Number of classes
Re0	Reuters-21578	1 792	12
Re1	Reuters-21578	12 363	87
Re2	www.reuters.com	464	6

Table 9
The six classes of Re2.

Sub class of entertainment	Number of documents	Number of words
Arts	30	1394
Film	122	3842
Music	70	2243
Industry	70	1885
Television	111	3062
People	61	1869

Table 10
Different weighting schemes.

Scheme	Acronym	Formula
Term frequency	TF	$tf(n)$
TF-inverse document frequency	TFIDF	$tf(n)\log df(n)$
Add-one log of TFIDF	LTF1IDF	$(1+\log tf(n)) * \log(1+m/df(n))$
TF-Discrimination Information	TF-DI	$tf(n) * DI(n)$

8. Document clustering based on DI

Current document clustering models generally include vector space document model [16,30], suffix tree document model [31–33] and document index graph model [34]. The common keyword weighting of these models is tf-idf; and the similarity between two documents is based on cosine measure. In this section, we employ a new weighting scheme based on DI of keywords for document clustering, and compare it with tf-idf weighting scheme experimentally.

8.1. Data sets

Three datasets are used in the experiment. The datasets Re0 and Re1 are from Reuters-21578 text categorization test collection. The dataset Re2 contains entertainment news from 1st/10/2008 to 31st/10/2008 downloaded from www.reuters.com. The summary of these datasets is given in Tables 8 and 9. In both of the datasets, we remove stop words such as "are", "is", and "in". Furthermore, since the DI is a measurement of noun, Stanford-Postagger is used to remove other words.

Besides, the basic K-means clustering technique [25] is used in our experiments. We evaluate the effectiveness of the document clustering with F-measure, Entropy, and Purity. Generally, we would like to maximize the F-measure and Purity, and minimize the Entropy of the clusters to achieve a high-quality document clustering [35].

8.2. The weighting scheme based on DI

Unlike tf-idf, tf-DI uses DI instead of idf (inverse document frequency) to measure the weight of the keywords, exploiting semantic discrimination capability based on semantics instead of inverse document frequency. The function is as follows:

$$wn = tf(n) * DI(n) \tag{15}$$

where $tf(n)$ is the frequency of the nth keyword in document d, and $DI(n)$ is DI of the nth keyword. Table 10 lists the four different weighting schemes.

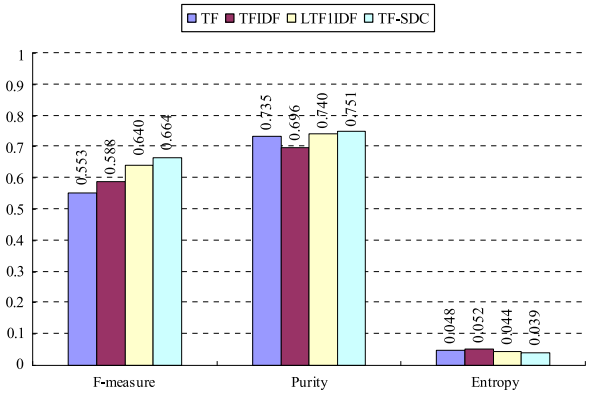

Fig. 6. The clustering results of Re0 data set.

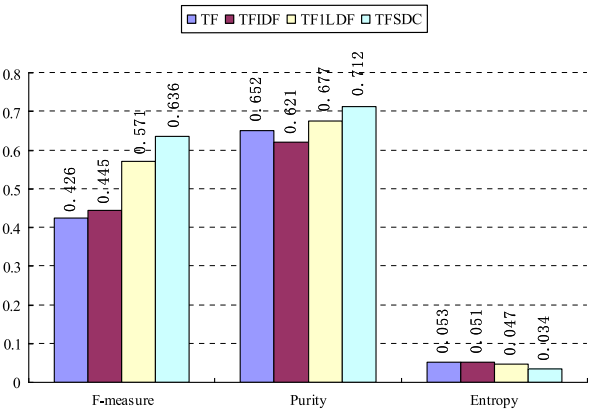

Fig. 7. The clustering results of Re0 data set.

8.3. Clustering results

Since the results of basic K-means for document clustering depend on the initial centroids, we carry out the basic K-means 100 times with randomly selected initial centroids to reduce the interference. Fig. 6 illustrates the average F-measure, Purity, and Entropy of 100 random initial centroids from Re0 dataset shown in Table 8. From Fig. 6, we can see that tf-DI performs the best among the four weighting schemes. The lift percentage of tf-DI against the other three schemes is low, which may be caused by the semantic correlation of classes in Re0 data set. Similar to the results of Re0 dataset, tf-DI also performs best in Re1 dataset which can be seen from Fig. 7. The documents of Re0 and Re1 are both get from Reuters-21578. The number of documents of Re1 is almost ten times than these of Re0. We can say that the better performance of tf-DI than other weighting scheme is irrelevant to the number of documents.

Fig. 8 illustrates the average F-measure, Purity, and Entropy of 100 random initial centroids from Re2 shown in Table 9 data set. From Fig. 8, we can see that tf-DI performs the best. The lift percentage of tf-DI against LTF1IDF is high. The reason for the high lift percentage may attribute to the high semantic correlation of the classes. For the data of Re2 downloaded from the entertainment news of www.reuters.com, the keyword overlaps in these classes are high, which may cause the high semantic correlation of the classes.

8.4. Discussions

(1) **tf-IDF vs. tf-DI**

From the experimental results in Figs. 6 and 7, it is apparent that tf-DI performs better than tf-IDF in document clustering,

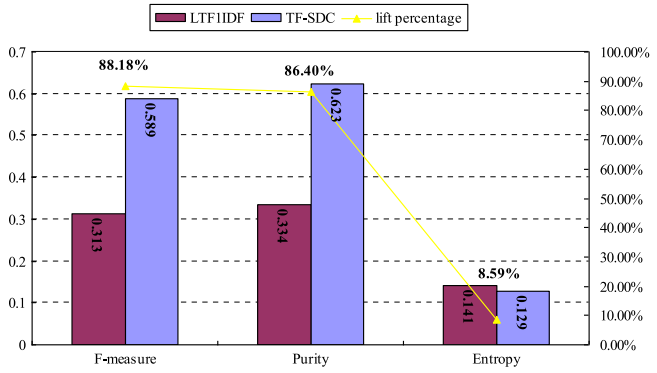

Fig. 8. The clustering results of Re1 data set.

as the average F-measure, Purity, and Entropy of tf-DI are higher than those of tf-IDF. The reason is due to that tf-DI uses semantic discrimination capability instead of inverse document frequency of keywords, which is more appropriate to discriminate documents. Besides the accuracy of tf-DI on document clustering, the complexity of computing DI is $O(nm)$, which is equal to the complexity of computing idf of keywords.

(2) **tf-DI vs. semantic correlation**

It is worth noting that tf-DI performs better on the Re2 dataset than on the Re0 and Re1 datasets. In practice, tf-DI performs better than tf-IDF in F-measure and Purity of all 100 random initial centroids on Re2 dataset. As for Re0 dataset, tf-DI performs better than tf-IDF in F-measure and Purity of a few 100 random initial centroids. The reason is due to the semantic correlation of Re2 being stronger than Re0 and Re1. In other words, the performance of tf-DI relies on the semantic correlation of classes in data sets. The stronger the semantic correlation between classes of data, the better tf-DI performs.

(3) **tf-DI vs. data mining tasks**

In [16], Salton et al. investigated the correlation between space density and indexing performance, and confirmed the usefulness of DI in information retrieval. In our work, we further demonstrate its utility in data mining tasks such as document clustering. A new weighting scheme of keywords based on DI is employed for document clustering, which shows that DI has a good performance on document clustering. Different from using space density as the measure of DI, our proposed model uses the feature of the document frequency spectrum and the power law function of keywords to compute DI. By employing the inherent feature of keywords as opposed to space density, experimental result show that DI has a good prospect in data mining area and IR applications.

9. Conclusion

In this paper, we have proposed a new model for computing discrimination information (DI). With respect to Salton et al.'s discrimination value model, the computing complexity of our model is $O(nm)$ for a collection of m documents with n keywords, which is in practice directly usable for a collection containing large documents. The proposed method does not need any prior knowledge such as the distribution of keywords, thus can be used on text mining related tasks easily. We make some contributions to the development of the measurement of discrimination information as follows.

(1) We defined three types of keywords (i.e. general keywords, functional keywords, and limited keywords) according to the document frequency spectrum of keywords, which leads us to a qualitative analysis of DI in a knowledge base system.

(2) We further used the power law function of keywords to identify the three types of keywords according to the difference of the document frequencies and the probability density function of keywords, serving as a footstone to reduce the computing complexity of DI.

(3) Based on the linear regression of the power law function of keywords, we developed an algorithm of computing DI, which transforms the problem of identifying the general keywords to an analysis of the distances between adjacent points in the linear regression line of keywords. Experimental results confirm that we can use a simpler algorithm to replace the complex model proposed by Salton et al. with a low error rate.

(4) In addition, we used the entropy to construct a piecewise function for computing DI, which transforms the discrimination capability of keywords into a scalable value that can be used directly. Meanwhile, four data sets are used to evaluate the accuracy of the qualitative analysis on DI. The high correlation coefficient between our model and Salton et al.'s model shows that our model can effectively compute DI with lower computing complexity.

(5) Finally, the accuracy of computation by our proposed model on the four data sets shows that it can make a good balance between the accuracy and the complexity of computing DI appropriately.

As a demonstration on the utility of our model, a new weighting scheme of keywords based on DI is employed for document clustering. Experimental results show that our model outperforms the term frequency-inverse document frequency (tf-idf) on document clustering by a wide margin. Moreover, the successful application on document clustering shows that DI has a good prospect on the information retrieval area.

References

[1] J. Kopecky, T. Vitvar, C. Bournez, J. Farrell, SAWSDL: Semantic annotations for WSDL and XML scheme, IEEE Internet Comput. 11 (6) (2007) 60–67.
[2] K. Nagao, Y. Shirai, K. Squire, Semantic annotation and transcoding: Making web content more accessible, IEEE Multimedia 8 (2) (2001) 69–81.
[3] C. Xu, J. Wang, H. Lu, Y. Zhang, A novel framework for semantic annotation and personalized retrieval of sports video, IEEE Trans. Multimedia 10 (3) (2008) 421–436.
[4] X. Luo, X. Wei, J. Zhang, Guided game-based learning using fuzzy cognitive maps learning technologies, IEEE Trans. Learn. Technol. 3 (4) (2010) 344–357.
[5] Y. Zhao, L. Feng, L. Chen, Detection of multi-relations based on semantic communities behaviors, in: International Conference on Service Systems and Service Management, 2007, pp. 1–7.
[6] H. Zhuge, X. Sun, P. Shi, Resource space model, OWL and database: Mapping and integration, ACM Trans. Internet Technol. 8 (4) (2008) 20–50.
[7] L. Gupta, S. Kota, S. Murali, L. Molfese, R. Vaidyanathan, A feature ranking strategy to facilitate multivariate signal classification, IEEE Trans. Syst. Man Cybern. Part C Appl. Rev. 40 (1) (2010) 98–108.
[8] S. Kim, K. Han, H. Rim, S. Myaeng, Some effective techniques for naive bayes text classification, IEEE Trans. Knowl. Data Eng. 18 (11) (2006) 1457–1466.
[9] O. Kurland, L. Lee, Clusters, language models, and ad hoc information retrieval, ACM Trans. Inf. Syst. 27 (3) (2009).
[10] E. Leung, Q. Li, An experimental study on personalized learning environment through open source software tools, IEEE Trans. Educ. 50 (4) (2007) 331–337.
[11] Q. Li, R. Lau, T. Shih, F. Li, Technology supports for distributed and collaborative learning over the Internet, ACM Trans. Internet Technol. 8 (2) (2008).
[12] Y. Gao, B. Zheng, G. Chen, Q. Li, C. Chen, G. Chen, Efficient mutual nearest neighbor query processing for moving object trajectories, Inform. Sci. 180 (11) (2010) 2176–2195.
[13] B. Smyth, A community-based approach to personalizing web search, Computer 40 (8) (2007) 42–50.
[14] R. Varadarajan, V. Hristidis, T. Li, Beyond single-page web search results, IEEE Trans. Knowl. Data Eng. 20 (3) (2008) 411–424.
[15] G. Xue, J. Han, Y. Yu, Q. Yang, User language model for collaborative personalized search, ACM Trans. Inf. Syst. 27 (3) (2009) 1–28.
[16] G. Salton, A. Wong, C. Yang, A vector space model for automatic indexing, Commun. ACM 18 (11) (1975) 613–620.
[17] D. Cai, An information theoretic foundation for the measurement of discrimination information, IEEE Trans. Knowl. Data Eng. 22 (9) (2010) 1262–1273.
[18] G. Salton, C. Yang, On the specification of keyword values in automatic indexing, J. Doc. 29 (4) (1973) 351–372.
[19] J. Leiva-Murillo, A. Artes-Rodriguez, Maximization of mutual information for supervised linear feature extraction, IEEE Trans. Neural Netw. 18 (5) (2007) 1433–1441.

Please cite this article in press as: Z. Xu, et al., Power law based foundation for the measurement of discrimination information for human knowledge representation, Future Generation Computer Systems (2016), http://dx.doi.org/10.1016/j.future.2016.10.021

[20] Y. Saygin, A. Reisman, Y. Wang, Value of information gained from data mining in the context of information sharing, IEEE Trans. Eng. Manage. 51 (4) (2004) 441–450.
[21] E. Wiener, J. Pedersen, A. Weigend, A neural network approach to topic spotting, in: Proceedings of the 4th Annual Symposium on Document Analysis and Information Retrieval, 1995, pp. 317–332.
[22] F. Topsoe, Some inequalities for information divergence and related measures of discrimination, IEEE Trans. Inform. Theory 46 (4) (2000) 1602–1609.
[23] A. Dembo, Information inequalities and concentration of measure, Ann. Probab. 25 (1997) 927–939.
[24] F. Österreicher, I. Vajda, Statistical information and discrimination, IEEE Trans. Inf. Theory 39 (3) (1993) 1036–1039.
[25] Y. Theoharis, Y. Tzitzikas, D. Kotzinos, V. Christophides, On graph features of semantic web schemes, IEEE Trans. Knowl. Data Eng. 20 (5) (2008) 692–702.
[26] J. Daintith, Oxford Dictionary of Physics, Oxford University Press, 2005.
[27] W. Press, S. Teukolsky, W. Vetterling, B. Flannery, Numerical Recipes in C, Cambridge University Press, 1992.
[28] G. Zipf, Human Behavior and the Principle of Least Effort, Addison-Wesley, 1949.
[29] P. Resnik, Using information content to evaluate semantic similarity in taxonomy, in: Proceedings of 14th International Joint Conference on Artificial Intelligence, 1995, pp. 448–453.
[30] M. Steinbach, G. Karypis, V. Kumar, A comparison of document clustering techniques, in: KDD Workshop on Text Mining, 2000.
[31] H. Chim, X. Deng, A new suffix tree similarity measure for document clustering, in: Proceedings of the 16th International Conference on World Wide Web, 2007, pp. 121–129.
[32] H. Chim, X. Deng, Efficient phrase-based document similarity for clustering, IEEE Trans. Knowl. Data Eng. 20 (9) (2008) 1217–1229.
[33] O. Zamir, O. Etzioni, Web document clustering: A feasibility demonstration, in: SIGIR'98, 1998.
[34] K.M. Hammouda, M.S. Kamel, Efficient phrase-based document indexing for web document clustering, IEEE Trans. Knowl. Data Eng. 16 (10) (2004) 1279–1296.
[35] S. Fodeh, B. Punch, P. Tan, On ontology-driven document clustering using core semantic features, Knowl. Inf. Syst. 28 (2011) 395–421.

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 as a postdoc at 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 Trans. on Systems, Man, and Cybernetics: Systems, etc.

Xiangfeng Luo is a professor in the School of Computers, Shanghai University, China. Currently, he is a visiting professor at Purdue University, USA. His main research interests include Web Wisdom, Cognitive Informatics, and Text Understanding. He has authored or co-authored more than 100 publications and his publications have appeared in IEEE Trans. on Systems, Man, and Cybernetics-Part C, IEEE Trans. on Automation Science and Engineering, IEEE Trans. on Learning Technology, etc. He has served as the Guest Editor of ACM Transactions on Intelligent Systems and Technology. Dr. Luo has also served on the committees of a number of conferences/workshops, including Program Co-chair of ICWL 2010 (Shanghai), WISM 2012 (Chengdu), CTUW2011 (Sydney) and PC member for more than 40 conferences and workshops.

Yunhuai Liu is a professor in the third research institute of ministry of public security, China. He received the Ph.D. degrees from Hong Kong University of Science and Technology (HKUST) in 2008. His main research interests include wireless sensor networks, pervasive computing, and wireless network. He has authored or co-authored more than 50 publications and his publications have appeared in IEEE Trans. on Parallel and Distributed Systems, IEEE Journal of Selected Areas in Communications, IEEE Trans. on Mobile Computing, IEEE Trans. on Vehicular Technology etc.

Lin Mei received his Ph.D. degree from Xi'an Jiaotong University, Xi'an, China, in 2000. He is a Research Fellow. From 2000 to 2006, he was a Postdoctoral Researcher with Fudan University, Shanghai, China; the University of Freiburg, Freiburg im Breisgau, Germany; and the German Research Center for Artificial Intelligence. He is currently the Director of the Technology R&D Center for the Internet of Things with the Third Research Institute of the Ministry of Public Security, China. He has published more than 40 papers. His research interests include computer vision, artificial intelligence, and big data processing.

Chuanping Hu received his Ph.D.degree from Tongji University, Shanghai, China, in 2007. He is a Research Fellow and the Director of the Third Research Institute of the Ministry of Public Security, China. He is also a specially appointed Professor and a Ph.D. supervisor with Shanghai Jiao Tong University, Shanghai, China. He has published more than 20 papers, has edited five books, and is the holder of more than 30 authorized patents. His research interests include machine learning, computer vision, and intelligent transportation systems. He is the chairman of ACM Shanghai Chapter.

Please cite this article in press as: Z. Xu, et al., Power law based foundation for the measurement of discrimination information for human knowledge representation, Future Generation Computer Systems (2016), http://dx.doi.org/10.1016/j.future.2016.10.021