Information Sciences 180 (2010) 4353–4369

Contents lists available at ScienceDirect

Information Sciences

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

Validation of overlapping clustering: A random clustering perspective

Junjie Wu [a,*], Hua Yuan [b], Hui Xiong [c], Guoqing Chen [d]

[a] School of Economics and Management, Beihang University, Beijing 100191, China
[b] School of Management and Economics, University of Electronic Science and Technology of China, Chengdu 610054, China
[c] Management Science and Information Systems Department, Rutgers University, Newark 07102, NJ, USA
[d] School of Economics and Management, Tsinghua University, Beijing 100084, China

ARTICLE INFO

Article history:
Received 13 January 2009
Received in revised form 20 July 2010
Accepted 26 July 2010

Keywords:
Information retrieval
Cluster validation
F-measure
Implication intensity (*IMI*)
Incomplete beta function

ABSTRACT

As a widely used clustering validation measure, the *F*-measure has received increased attention in the field of information retrieval. In this paper, we reveal that the *F*-measure can lead to biased views as to results of overlapped clusters when it is used for validating the data with different cluster numbers (incremental effect) or different prior probabilities of relevant documents (prior-probability effect). We propose a new "IMplication Intensity" (*IMI*) measure which is based on the *F*-measure and is developed from a random clustering perspective. In addition, we carefully investigate the properties of *IMI*. Finally, experimental results on real-world data sets show that *IMI* significantly alleviates biased incremental and prior-probability effects which are inherent to the *F*-measure.

1. Introduction

The Internet provides a vast resource of information and services that are continuing to grow rapidly. Powerful search engines have been developed to aid in locating relevant documents by categories, contents, or subjects [29]. While tremendous efforts have been made on improving the performances of these search engines, the results returned by the search engines still contain many documents that meet the search criteria but are of no interest to the users. This has led to an increased interest in developing methods that can help users effectively navigate, summarize, and organize search results.

Efficient and effective text clustering algorithms are now playing an important role in optimizing search performance [42]. Text clustering provides insight into the documents by dividing the documents into groups (clusters) in a way such that documents in the same cluster are more similar to each other than documents in different clusters. Text clustering has been shown to provide an intuitive navigation mechanism by organizing large amounts of information into a small number of meaningful clusters as well as to greatly improve the retrieval performance by cluster-driven dimensionality reduction [14], term-weighting, and/or query expansion techniques [41].

Cluster validation is an important step in the clustering process. A common way of cluster validation is to use objective and quantitative validation measures [11]. In the literature, there have been considerable research efforts on designing and studying the cluster validation measures [28,11,7,8,25,33,2,20]. Among these available measures, the *F*-measure is widely used for cluster validation in document clustering [13,37,34,18]. There are also some other external measures, such as Entropy, Mutual Information, Variation of Information, Rand index, and Γ statistic [19,32].

Mehlitz et al. [18] have studied a limitation of the *F*-measure, which in this paper is denoted as the "incremental effect" of the *F*-measure. In other words, the *F*-measure tends to assign higher scores to the clustering results containing a large

* Corresponding author.
E-mail address: wujj@buaa.edu.cn (J. Wu).

0020-0255/$ - see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.ins.2010.07.028

number of clusters. This is a latent bias and is not easy to be perceived by the users. In addition, we also find that the *F*-measure has the "prior-probability effect" when it is used for data sets with different prior distributions. Specifically, the *F*-measure tends to assign higher scores to the clustering results with higher prior probabilities for the relevant documents. Indeed, from a random clustering perspective, these two effects reflect the systematic errors of the *F*-measure and may lead to a biased validation on the clustering results.

To meet the above critical challenges, we propose a novel "IMplication Intensity" (*IMI*) measure which is developed on top of the *F*-measure from a random clustering perspective. *IMI* is essentially a probability measure that shows how well a clustering result matches the result of random clustering. Empirical studies show that *IMI* is computationally efficient and demonstrate why *IMI* is capable of handling both the incremental and prior-probability effects of the *F*-measure.

Finally, we have conducted extensive experiments on real-world document data sets. Results show that, as an enhanced *F*-measure, *IMI* well alleviates the incremental and prior-probability effects of the *F*-measure. Also, *IMI* provides better discrimination of clustering results than the *F*-measure.

The remainder of this paper is organized as follows. Section 2 presents document clustering algorithms and the corresponding cluster validation measures. In Section 3, we illustrate the incremental effect and the prior-probability effect inherent to the *F*-measure. The notion of the implication intensity measure is introduced in Section 4. Section 5 gives the experimental results. Finally, we conclude our work in Section 6.

2. Preliminaries

In this section, we review some main issues related to cluster validation in information retrieval. Specifically, we briefly introduce the text clustering algorithms and the cluster validation measures used in this paper. Table 1 shows some mathematical notation used throughout the paper.

2.1. Text clustering algorithms for information retrieval

Information Retrieval (IR) is the science of searching for relevant information [31,16]. And text clustering [3,35,27] has played an important role in many applications of information retrieval.

A text clustering algorithm partitions a set of documents in a way such that documents in the same group have similar contents and documents in different groups tend to have different contents. Text clustering has been applied to the documents retrieved by a search engine so that the information contained in these articles can be presented more effectively to the users [12,9,41]. Next, we introduce the Frequent Term Based Clustering (FTC), a text clustering algorithm used in this paper.

The *Frequent Term Based Clustering (FTC)* algorithm [5] is based on the observation that frequent term sets [1] in the documents are the key for text clustering. In other words, different text clusters are formed around different frequent term sets. Along this line, the procedure of FTC can be divided into two phases. In the first phase, FTC discovers all the frequent term sets by the well-established association mining algorithms such as Apriori [1] and FP-tree [10]. In the second phase, the documents which contain all the terms in a frequent set are grouped into one cluster. Since one document may contain all the terms of several frequent sets, it is natural to see that the resulting clusters overlap. The problem is, however, that this overlapping is usually too heavy due to the "downward closure" property of the frequent sets. To cope with this, FTC proposed a so-called "Entropy Overlap" (*EO*) measure to rank the frequent term sets. Let C_i denote cluster i, d_j denote the retrieved document j, f_j denote the number of frequent term sets supported by document j, then the *EO* of cluster C_i can be computed as follows:

$$EO(C_i) = \sum_{d_j \in C_i} -\frac{1}{f_j} \ln\left(\frac{1}{f_j}\right). \tag{1}$$

Table 1
Mathematical notation.

D	The entire document set returned by a search engine
T	The number of documents in D, i.e., $T = \|D\|$
P	The number of relevant documents in D
c	The number of clusters based on which we evaluate clustering performances
t_i	The number of documents in cluster i, $i = 1,\ldots,c$
p_i	The number of relevant documents in cluster i, $i = 1,\ldots,c$
p^*	The maximum number of relevant documents among c clusters, i.e., $p^* = \max_{i=1}^{c} p_i$
$F(i)$	The F-measure score for cluster i, $i = 1,\ldots,c$
F^*	The maximum F-measure score among c clusters, i.e., $F^* = \max_{i=1}^{c} F(i)$

A lower *EO* value indicates higher rank. This means frequent sets with lower *EO* values have the priority to form clusters, and the frequent sets with higher *EO* values may be discarded if all the documents have been covered by higher-ranked frequent sets at least once.

FTC has the merit of high efficiency and can generate overlapped clusters. These two characteristics are crucial for the successful use of text clustering in information retrieval since (1) the documents to be clustered are often in huge volume with high dimension, and (2) it is natural to read a document containing topics in various domains. There are many other well-established text clustering algorithms such as Scatter/Gather [6], SuffixTree Clustering [40], and bisecting *K*-means [24].

2.2. Cluster validation measures for text clustering

In information retrieval, a common validation method for text clustering is the so-called "Optimal Cluster Validation" (*OCV*) strategy [26,17]; that is, the validation system first specifies some cluster validation measures as the validity criterion, then searches the clusters produced by a clustering tool to find the one with the highest validation score, and finally returns this score as the validation of the whole clustering. Since *OCV* has been widely adopted in evaluating information retrieval systems, we restrict our study to *OCV*.

Many different measures that can be used in *OCV* to evaluate cluster qualities have been proposed. Most of them assume a ground truth notion of relevancy: Every document is known to be either relevant or irrelevant to a particular query. Given the notations in Table 1, in what follows, we list some of the representative measures we intent to study in this paper.

2.2.1. Recall and precision

These are two of the most widely used measures for information retrieval [25]. Recall (*Rec*) is the fraction of the retrieved documents that are relevant to the query. That is

$$Rec(i) = \frac{p_i}{P}, \quad 1 \leqslant i \leqslant c. \tag{2}$$

It is trivial to achieve a recall of 100% by returning all documents in response to any query. Therefore, recall alone is not enough but one needs to measure the number of non-relevant documents also, for example, by computing the precision (*Prec*) as follows:

$$Prec(i) = \frac{p_i}{t_i}, \quad 1 \leqslant i \leqslant c. \tag{3}$$

Apparently, precision is the fraction of the documents retrieved within a cluster that are relevant to the user's information need.

2.2.2. F-measure

F-measure [13] is the weighted harmonic mean of precision and recall, which provides a reasonable way to integrate the validation of these two measures, and tends to highlight the smaller one. Typically, we have

$$F(i) = \frac{2}{\frac{1}{Prec(i)} + \frac{1}{Rec(i)}} = \frac{2p_i}{t_i + P}, \quad 1 \leqslant i \leqslant c. \tag{4}$$

This is also known as the F_1 measure, because its recall and precision are evenly weighted. The general formula for the *F*-measure with the non-negative real β parameter is:

$$F_\beta(i) = \frac{(1+\beta^2)Prec(i)Rec(i)}{\beta^2 Prec(i) + Rec(i)} = \frac{(1+\beta^2)p_i}{t_i + \beta^2 P}, \quad 1 \leqslant i \leqslant c. \tag{5}$$

F_β was derived by van Rijsbergen [28] which "measures the effectiveness of retrieval with respect to an user who attaches β times as much importance to recall as precision". Besides F_1, people often use $F_{0.5}$ and F_2 in which $\beta = 0.5$ and 2, respectively. Also, it is noteworthy that F_β is in the range of [0,1].

In summary, the recall, precision and *F*-measure are the widely used cluster validation measures we focus in this paper. To avoid complexity, we hereby agree that if there is no confusion, any measure we mentioned in this paper is used with the *OCV* scheme. For instance, "the problem of the *F*-measure" means "the problem of the *F*-measure in the *OCV* scheme".

3. Issues with cluster validation measures

In this section, we present some issues with the cluster validation measures in the *OCV* scheme. We first illustrate it by a random clustering example.

3.1. The incremental effect of the F-measure

For an application scenario of information retrieval, the search engine returns *T* documents for the query, among which *P* documents are indeed relevant. Then we perform multiple random clusterings on the *T* documents. That is, we indepen-

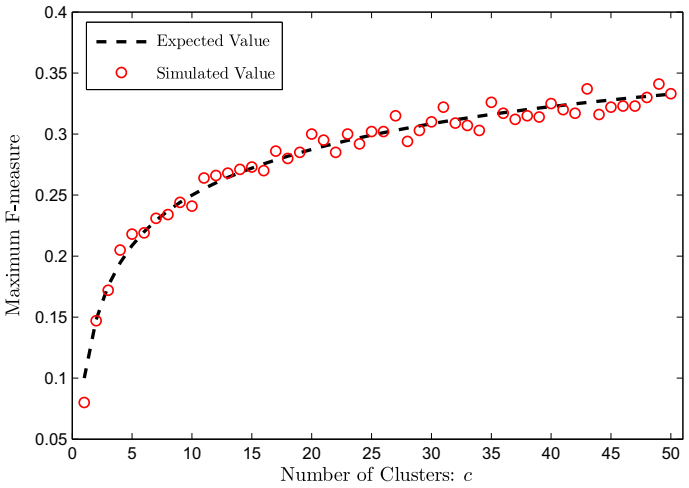

Fig. 1. Maximum *F*-measure versus *c*. ($T = 100$, $P = 10$, $\tau = \{t_i = 10 | 1 \leqslant i \leqslant c\}$).

dently repeat the follow procedure *c* times: Randomly draw without replacements *t* documents from the returned *T* documents, and form a cluster by these *t* documents. By doing so, we finally have *c* clusters that may be overlapped and with the same size *t*. We denote the above random clustering as $\mathcal{RC}(T, P, c, \tau)$, where $\tau = \{t_i = t | 1 \leqslant i \leqslant c\}$. Next, we exploit *OCV* with *F*-measure to evaluate the results by random clusterings. Fig. 1 shows the results. Note that for each *c*, we repeated $\mathcal{RC}(T, P, c, \tau)$ 100 times and returned the averaged *F*-measure values.

As indicated by the circle markers (simulated values) in Fig. 1, if we use the *F*-measure to evaluate the random clusterings, the *F*-measure value tends to increase with the increase of the number of clusters *c*. This implies that for two clustering results, given other conditions are the same, the one with more clusters tends to be assigned with a higher score by the *OCV* system. In other words, *OCV* with *F*-measure tends to "favor" clustering results containing more clusters, which is a systematic but latent feature that could be misleading and not easy to be aware of by the users.

Actually in [17,18], Mehlitz et al. presented a detailed mathematical description of how to compute the expected maximum number of relevant documents within each cluster produced by the above random clustering scheme. We reorganize it into the following proposition:

Proposition 1. *Given a random clustering* $\mathcal{RC}(T, P, c, \tau)$ *with* $\tau = \{t_i = t | 1 \leqslant i \leqslant c\}$, *the expected maximum number of relevant documents among the c clusters is*

$$E(p^*) = \sum_{j=0}^{t} j \frac{\sum_{i=1}^{c} \binom{c}{i} \left(\binom{P}{j} \binom{T-P}{t-j} \right)^i \left(\sum_{k=0}^{j-1} \binom{P}{k} \binom{T-P}{t-k} \right)^{c-i}}{\binom{T}{t}^c}. \tag{6}$$

We leave the proof to Appendix A. According to Proposition 1 and Eq. (4), since *t* and *P* are two constants, the expected Maximum *F*-measure is

$$E(F^*) = \frac{2E(p^*)}{t + P}.$$

Thus we can compute a series of $E(F^*)$ values given different *c* values, as shown by the black dash line in Fig. 1. As can be seen, the expected values well match the simulated values, which further justifies our analysis: *OCV* with *F*-measure tends to favor clustering results with more clusters.

The same problem actually exists for *OCV* with recall, precision, and F_β, since we can easily have

$$E(Rec^*) = \frac{E(p^*)}{P},$$
$$E(Prec^*) = \frac{E(p^*)}{t},$$
$$E\left(F_\beta^*\right) = \frac{(1 + \beta^2)E(p^*)}{\beta^2 P + t},$$

where Rec^*, $Prec^*$ and F_β^* are the maximum recall, precision and F_β values among the clusters generated by the random clusterings, respectively. This implies that it is the *OCV* mechanism rather than the measures that brings systematic errors into the validation process. Indeed, it is not difficult to understand that as the number of clusters increases, we have more

Table 2
Two clusterings: an example.

	T	P	t^a	p^a	F^a
$\mathcal{D}_1$	1000	800	400	320	0.53
$\mathcal{D}_2$	1000	200	400	120	0.40

[a] Means the value of the optimal cluster.

chances to get a cluster with a higher score if *OCV* is used. Hereinafter we call it the "incremental effect" of the *F*-measure in *OCV*, or simply the incremental effect of the *F*-measure.

3.2. The prior-probability effect of the F-measure

We further explore whether the prior distributions of the relevant documents can have impact on the cluster validation. We illustrate it by a simulated example as follows.

Suppose we have two collections of the retrieved documents: $\mathcal{D}_1$ and $\mathcal{D}_2$. Details of these two data sets and the optimal clusters produced by some clustering tool are listed in Table 2.

As indicated by Table 2, $\mathcal{D}_1$ and $\mathcal{D}_2$ both contain 1000 documents, but differ in the number of relevant documents – $\mathcal{D}_1$ contains 800 relevant documents but $\mathcal{D}_2$ contains only 200 relevant documents. That is to say, the prior-probability of the relevant documents in $\mathcal{D}_1$ is $P_1/T_1 = 0.8$, much larger than the one in $\mathcal{D}_2 : P_2/T_2 = 0.2$. Now we assume after clustering, the optimal cluster sizes for the two data sets are the same as 400, and the numbers of relevant documents in the optimal clusters are 320 and 120, respectively. What is the validation by the *F*-measure?.

It is not difficult to compute that the *F*-measure of the clustering on $\mathcal{D}_1$ is 0.53, much larger than 0.40, the one of the clustering on $\mathcal{D}_2$, as shown in Table 2. This implies that the *F*-measure tends to favor the first clustering. However, if we compute the precisions of the two clusterings, we can find that $Prec_1 = 0.8$, a value equal to the prior-probability (0.8), whereas $Prec_2 = 0.3$, a value higher than the prior-probability (0.2). Thus from a statistical viewpoint, the performance of the first clustering is merely comparable to the performance of a random clustering, and the second clustering is significantly better than the random clustering. Also note that the second clustering has a higher recall value than the first one. Therefore, from a statistical point of view, we cannot conclude that the first clustering is better than the second one.

In summary, the *F*-measure tends to favor the clustering with a higher prior-probability for the relevant documents. Hereinafter, we call it the "prior-probability effect" of the *F*-measure.

3.3. Problem formulation

Here, we formulate the problem as follows:

3.3.1. Problem definition

Design a cluster validation scheme that can address the incremental effect and the prior-probability effect inherent to the *F*-measure when validating clustering results in information retrieval.

4. Random clustering validation

In this section, we introduce a novel "Random Clustering Validation" (*RCV*) scheme based on the "implication intensity" measure for text clustering. And we further enhance *RCV* by combining implication intensity with the *F*-measure to form a "Mixed Validation" (*MV*) scheme.

4.1. The concept of matched random clustering

Definition 1 (*Matched Random Clustering,* $\mathcal{MRC}$). Let $\mathcal{C}(D, T, P, c, \tau)$ denote a clustering on a document data set D with T documents and P relevant documents, where the number of clusters is c and $\tau = \{t_i | 1 \leqslant i \leqslant c\}$ represents the cluster sizes. A random clustering $\mathcal{RC}(T', P', c', \tau')$ is called the matched random clustering of $\mathcal{C}(D, T, P, c, \tau)$ if and only if: (I) $T = T'$, (II) $P = P'$, (III) $c = c'$, and (IV) $\tau = \tau'$.

Remark. As it is known, a text clustering relies heavily on the attribute values of data sets, but its $\mathcal{MRC}$ is attribute irrelevant; that is, the four parameters T, P, c and τ exclusively decide an unique $\mathcal{MRC}$.

Next, we discuss the function of $\mathcal{MRC}$ in measuring the clustering performance of $\mathcal{C}$. Actually, most of us can have the intuition that the clustering results produced by $\mathcal{MRC}$ should be much poorer than the ones produced by $\mathcal{C}$, since $\mathcal{C}$ is often designed purposefully by researchers to fulfill the document clustering tasks, but $\mathcal{MRC}$ is simply based on a random mechanism. That is to say, the clustering performance of the "virtual" $\mathcal{MRC}$ can serve as the baseline for the measuring of the "real" $\mathcal{C}$. This leads to our "random clustering validation" scheme described below.

4.2. Random clustering validation based on $\mathcal{MRC}$

Now we illustrate how to use $\mathcal{MRC}$ to derive the so-called "Random Clustering Validation" (*RCV*) for text clustering.

Assume that the clustering $\mathcal{C}(D,T,P,c,\tau)$ has an *F*-measure value F^*, and its matched random clustering $\mathcal{MRC}(T,P,c,\tau)$ has an *F*-measure value f^*. Let us consider the probability that the performance of $\mathcal{MRC}$ is poorer than the performance of $\mathcal{C}$, i.e., $\Pr(f^* < F^*)$. From a statistical point of view, a higher value of this probability implies a better clustering quality of $\mathcal{C}$ "tuned" by its $\mathcal{MRC}$. We call this probability the "*IMplication Intensity*", denoted by *IMI*. In what follows, we describe how to compute the implication intensity measure.

Proposition 2. *Given a clustering $\mathcal{C}(D,T,P,c,\tau)$ with the F-measure score F^*, and its $\mathcal{MRC}$ with the F-measure score f^*, let $I_x(a,b)$ denote the regularized incomplete beta function, we have*

$$IMI = \Pr(f^* < F^*) \approx \prod_{i=1}^{c}(1 - I_x(a_i, b_i)), \tag{7}$$

where

$$x = P/T,$$
$$a_i = \lceil 0.5F^*(P + t_i)\rceil,$$
$$b_i = t_i - a_i + 1.$$

We leave the proof to Appendix B. It is natural to extend Proposition 2 to a more general F_β case, as shown below.

Proposition 3. *Given a clustering $\mathcal{C}(D,T,P,c,\tau)$ with the F_β score S^*, and its $\mathcal{MRC}$ with the F_β score s^*, let $I_x(a,b)$ denote the regularized incomplete beta function, we have*

$$IMI = \Pr(s^* < S^*) \approx \prod_{i=1}^{c}(1 - I_x(a_i, b_i)), \tag{8}$$

where

$$x = P/T,$$
$$a_i = \left\lceil \frac{S^*\left(\beta^2 P + t_i\right)}{1+\beta^2} \right\rceil,$$
$$b_i = t_i - a_i + 1.$$

Since the proof is similar to the proof of Proposition 2, we omit it here. Now based on the implication intensity measure, we formulate our random clustering validation scheme as follows:

Definition 2 (*Random Clustering Validation, RCV*). For a text clustering $\mathcal{C}(D,T,P,c,\tau)$, a random clustering validation takes the following procedures:

(1) Computing the validation score by *OCV* with the *F*-measure.
(2) Computing the implication intensity value by Eq. (8).
(3) Returning the *IMI* value as the adjusted validation score.

A higher *IMI* value indicates a better outcome of clustering. Indeed, in contrast to the *F*-measure in *OCV*, the implication intensity in *RCV* is suitable for the comparison of clustering results with different cluster numbers, since it has been statistically adjusted by the $\mathcal{MRC}$ of the clustering.

4.3. Properties of the IMI measure

Here, we would like to explore the precision of the approximations used in deriving the *IMI* measure, and study the impact of the number of clusters and the prior-probability of relevant documents on the *IMI* measure.

4.3.1. The precision of the approximations

As formulated in Eq. (B.7), the first approximation in deriving the *IMI* measure is to replace the sampling without replacement by the sampling with replacement, so that we can use the binomial distribution results.

To know the precision of the approximation, we exploit two sampling strategies to compute the accumulated density function values of p_i (the number of relevant documents in cluster i), as shown in Fig. 2. As can be seen, when T is small relative to t_i, the difference between the two sampling strategies is obvious. However, as T increases, the gap narrows gradually. Consider that in real-world information retrieval practices, T is often much larger than t_i, this approximation is considered quite acceptable.

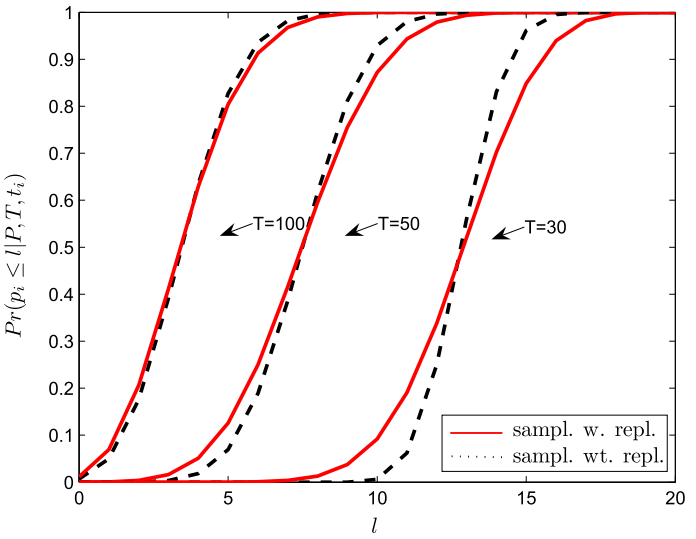

Fig. 2. "Sampling with replacements" versus "sampling without replacements." ($P = 20$, $t_i = 20$).

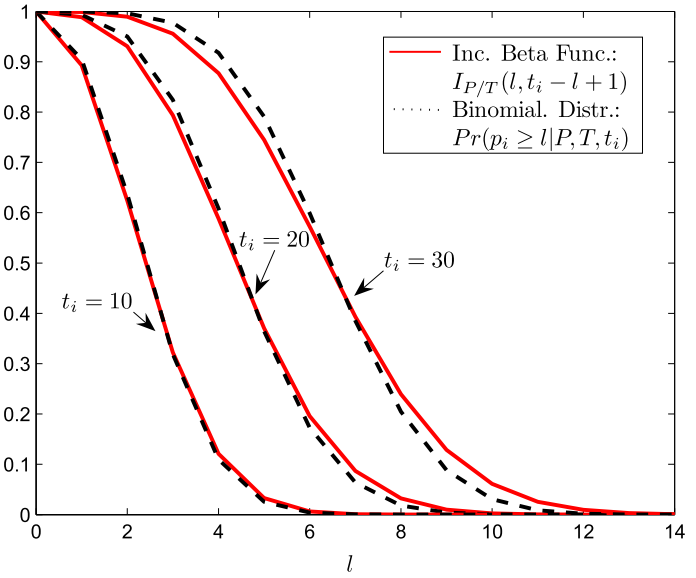

Fig. 3. "Incomplete beta function" versus "binomial distribution." ($T = 100$, $P = 20$).

The second approximation is the use of the regularized incomplete Beta function instead of the binomial distribution computation in Eq. (B.13). Since we approximate $Pr(p_i \geqslant l|P,T,t_i)$ by $I_{P/T}(l, t_i - l + 1)$, we compare these two quantities along different l values in Fig. 3. As can be seen, there are two observations: (1) Generally speaking, the difference between the two quantities is at an acceptable level, and (2) t_i serves as a key factor for the scale of the difference. In other words, a higher t_i value tends to increase the gap. Therefore, a not-so-large t_i value is beneficial for the precision of the second approximation.

In summary, the two approximations in deriving the implication intensity measure are considered reasonable, since the number of total documents (T) is often much larger than the cluster size (t_i) in real-world applications.

4.3.2. The effect of the cluster number

According to Eq. (8), five parameters, namely c, F^*, P/T, P and τ, can have impacts on the implication intensity measure. Among them, the parameters F^* and c are our focuses. We thus exploit sensitivity analysis for the two parameters by fixing the other parameters presented in Eq. (8).

As can be seen in Fig. 4, the implication intensity value increases as the increase of the F-measure value. This certainly agrees with our intuition, and it implies that the implication intensity has "inherited" the basic cluster validation ability

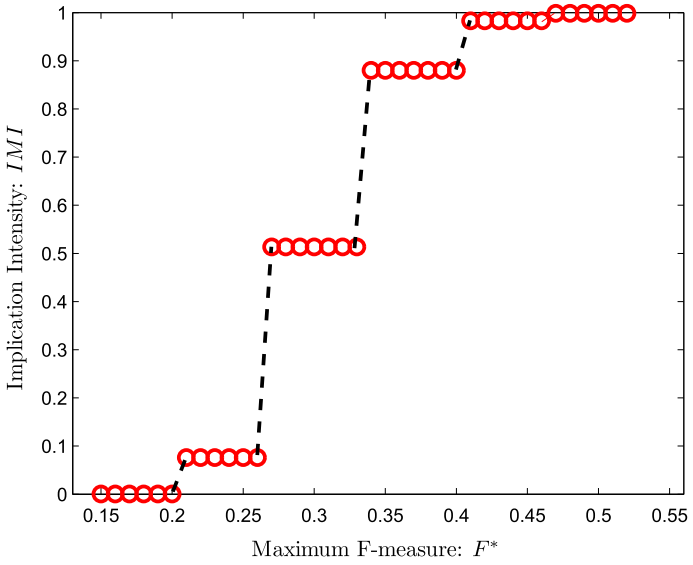

Fig. 4. Implication intensity versus F^*. ($0.15 \leqslant F^* \leqslant 0.52$, $c = 20$, $P/T = 0.2$, $P = 20$, $\{t_i = 10 | 1 \leqslant i \leqslant c\}$).

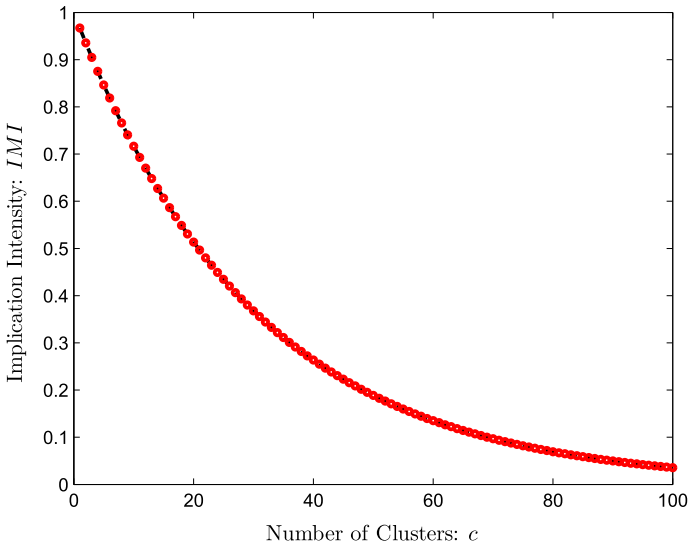

Fig. 5. Implication intensity versus c. ($1 \leqslant c \leqslant 100$, $F^* = 0.3$, $P/T = 0.2$, $P = 20$, $\{t_i = 10 | 1 \leqslant i \leqslant c\}$).

of the *F*-measure. More importantly, however, there exists a significant difference between the implication intensity and the *F*-measure; that is, the implication intensity value decreases as the increase of the number of clusters, as shown in Fig. 5. In other words, the implication intensity tends to "penalize" the clustering results involving more clusters, given the other parameters are the same. This is reasonable, since as the increase of the number of clusters, we tend to pay a higher search cost to find the documents we really have interests in.

Now let us put together all the impacts introduced by the number of clusters to the implication intensity in real-world scenarios. According to Figs. 1 and 5, the increase of c results in two effects: one is the increase of the *F*-measure value, and the other is the decrease of the implication intensity value. As indicated by Fig. 4, however, the first effect also has a "side effect" which tends to increase the implication intensity value. As a result, whether the increase of c can increase or decrease the value of the implication intensity is determined by the real performances of the clustering tools. That is to say, the use of the implication intensity measure avoids the systematic error brought by the different cluster numbers when using the *F*-measure to compare different clustering results.

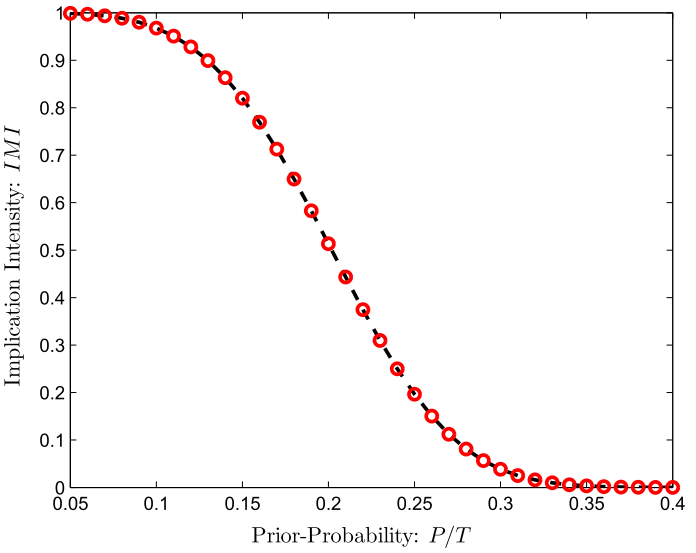

Fig. 6. Implication intensity versus P/T. ($0.05 \leqslant P/T \leqslant 0.40$, $c = 20$, $F^* = 0.3$, $P = 20$, $\{t_i = 10 | 1 \leqslant i \leqslant c\}$).

4.3.3. The effect of the prior-probability

To study the impact of the prior-probability of relevant documents (P/T) on the implication intensity measure, we also exploit sensitivity analysis.

As indicated by Fig. 6, if other conditions are the same, the implication intensity value decreases with the increase of the prior probability of the relevant documents. This is quite reasonable, since as the prior-probability increases, we would expect to see the optimal cluster with a better but not the same clustering quality. This implies that the use of the implication intensity measure can avoid the prior-probability effect of the F-measure when comparing different clustering results.

4.3.4. The comparison with the E_{abs} measure

As mentioned in Section 3.1, this work is inspired by Mehlitz et al. [18] in which the authors proposed a new measure called E_{abs} to handle the incremental effect of F-measure. In what follows, we study some properties of E_{abs} and compare E_{abs} with *IMI*.

As indicated by Mehlitz et al. [18], E_{abs} can be computed as follows:

$$E_{abs} = 1 - \frac{2P_{abs}R_{abs}}{P_{abs} + R_{abs}} \tag{9}$$

with

$$P_{abs} = (p^* - E(p^*))/\bar{t}, \quad R_{abs} = (p^* - E(p^*))/P, \tag{10}$$

where $E(p^*)$ can be computed by Eq. (6). Note that since we assume that all the clusters have the same size t in Eq. (6), $\bar{t}$ here is the averaged size of the clusters returned. Also, it is noteworthy that E_{abs} is a negative measure; that is, a higher E_{abs} value indicates a worse clustering result.

Next, we explore the relationships between E_{abs} and the parameters c and P/T. Results are shown in Fig. 7(a) and (b), respectively. Note that the parameters used for the two figures are the same as Figs. 5 and 6, respectively. As can be seen in the figures, E_{abs} shows similar properties as *IMI*; that is, given other conditions unchanged, E_{abs} increases with the number of clusters c and the prior-probability of relevant documents P/T, respectively. In other words, for the simulation cases, E_{abs} shows the ability to handle the incremental effect and the prior-probability effect of the F-measure.

Nevertheless, E_{abs} differs from *IMI* in various aspects as follows:

- *The value range problem*: E_{abs} has a much narrower range than *IMI*, and its upper bound can be greater than 1, as indicated by Fig. 7(a) and (b). This implies that E_{abs} has a poorer discrimination than *IMI*, and it cannot be used as a normalized measure (in the range of $[0, 1]$) to compare the clustering results of different document sets.
- *The cluster size puzzle*: One major problem of E_{abs} for real-world applications is the existence of unequal cluster sizes. The computation of $E(p^*)$ in Eq. (6) requires a same size for all the clusters returned, which is fine for the simulation cases in Fig. 7(a) and (b), but may encounter a great problem for real-world cases with clusters in varied sizes. We will detail this in the experimental section.
- *A combinatorial problem*: Another major problem of E_{abs} arises from the computations of the combinative numbers in Eq. (6). For instance, even for a small scale document size, say $T = 100$ in the above simulation cases, the computation of $\binom{T}{t}$

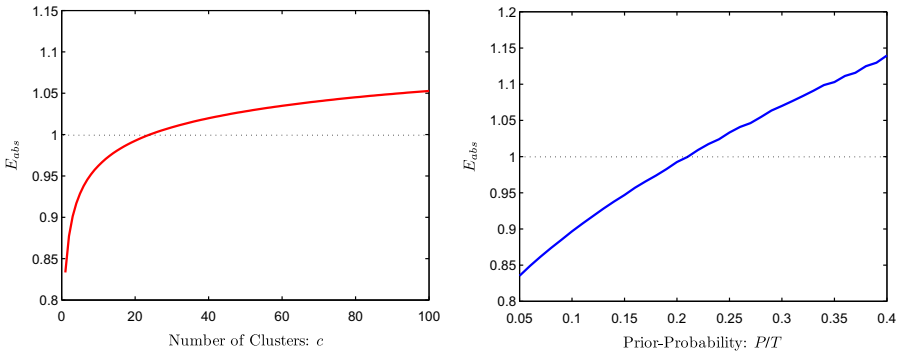

Fig. 7. Some properties of E_{abs}.

may often be imprecise. For real-world scenarios with large scale documents, the computation tends to be even infeasible. More details can be found in the experimental section.

4.4. Mixed validation

Every measure has its own limitations, so does the *IMI* measure. Indeed, the value range of *IMI* is not very wide. For instance, let us take two parameters F^* and P/T. In Fig. 4, for the parameters in the caption of this figure, the value of *IMI* is 0 when $F^* \leqslant 0.20$, and 1 when $F^* \geqslant 0.47$. In other words, if two clusterings have the F^* values below 0.2 or above 0.47, *IMI* cannot tell a difference in their clustering performances. A similar case holds for the P/T parameter as shown in Fig. 6. For the parameters used, *IMI* cannot make a distinction between the two clusterings with P/T values greater than 0.35.

The above leads to the so-called "Mixed Validation" (*MV*) scheme as follows:

Definition 3 (*Mixed Validation, MV*). For a text clustering $\mathcal{C}(D, T, P, c, \tau)$, the mixed validation takes the following procedures:

(1) Computing the validation score F^* by *OCV* with the *F*-measure.
(2) Computing the implication intensity value *I*.
(3) Returning *I* as the primary score, and F^* as the secondary score.

Remark. According to the *MV* scheme, to compare two clusterings, we first use *IMI* as the criterion; that is, the one with a higher *IMI* value is better. If the two clusterings have the same *IMI* value, we use the *F*-measure as the secondary criterion. In other words, the one with a higher *F*-measure value is better. If these two criteria yet cannot discriminate the two clusterings, we conclude that their clustering performance is not distinguishable with these two clustering validation measures.

5. Experimental results

In this section, we demonstrate the incremental effect and the prior-probability effect of the *F*-measure, and show how the implication intensity measure can address these issues.

5.1. Experimental tools

A number of existing text clustering tools have been used in the experiments. Specifically, we employ the simple FTC algorithm as described in the preliminary section except that frequent patterns (term sets) are replaced by hyperclique patterns [38,36]. This is due to the reason that the Apriori algorithm usually generates too many nested frequent patterns which will result in heavily overlapped clusters in the second phase of FTC. By contrast, hyperclique patterns do not contain cross-support patterns and are suitable for the subsequent clustering tasks. For more details about hyperclique patterns, please refer to [38,36].

5.2. Experimental data

Three real-world document data sets have been used in the experiments. The data set `re0` is from the Reuters-21578 text categorization collection Distribution 1.0 [15]. For `re0`, we selected documents that have a single label. The `la1` data set was obtained from the articles of the Los Angeles Times collected in TREC-5 [27]. The categories correspond to the desk of the

Table 3
Experimental data sets.

Data set	Source	#objects	#features	#classes	Min class size	Max class size
re0	Reuters-21578	1504	2886	13	11	608
la1	TREC-5	3204	21,604	6	273	943
wap	WebACE	1560	8460	20	5	341

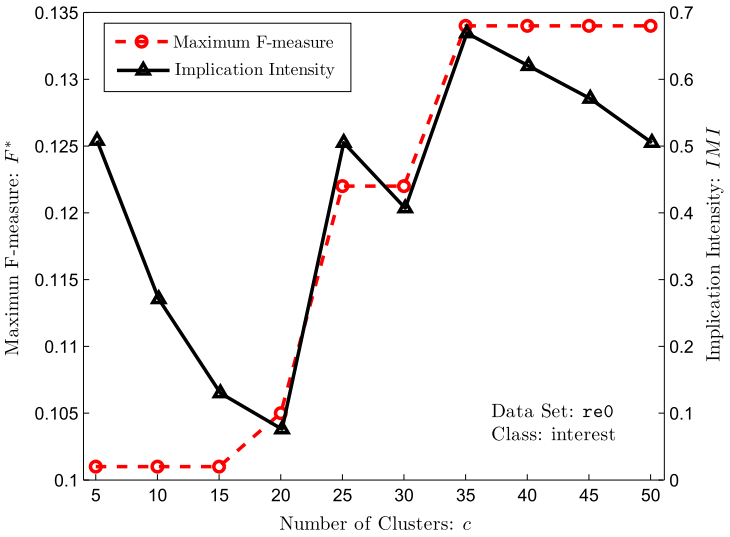

Fig. 8. *IMI* versus maximum *F*-measure: the incremental effect. (Parameters for hyperclique patterns: minimum support – 0.05, minimum *h*-confidence – 0.8).

paper that each article appeared and include documents from the entertainment, financial, foreign, metro, national, and sports desks. Finally, the data set `wap` was from the WebACE project [9]. Each document corresponds to a web page listed in the subject hierarchy of Yahoo! [39]. For the three data sets, we used a stop-list to remove common words, and the words were stemmed using Porter's suffix-stripping algorithm [21]. Note that these three are all multi-class data sets. From an information retrieval viewpoint, we can view them as the retrieved texts by some search engines, and take the documents from any class as the relevant documents we want. Some characteristics of the data sets are listed in Table 3.

5.3. The incremental effect

Here, we first illustrate the incremental effect of the *F*-measure. To this end, we generate hyperclique patterns from the data set `re0`, and then use FTC to produce overlapped clusters based on the identified hyperclique patterns. Fig. 8 shows the cluster validation results by having class `interest` as the relevant class.

As indicated by the dot line in Fig. 8, if we use the *F*-measure as the validation criterion, the score tends to be higher as the increase of the number of clusters returned by FTC. This is the incremental effect introduced by the *OCV* scheme. The negative impact of the incremental effect lies in two aspects. On one hand, it hinders us from comparing the clustering performances in an objective way, since the clustering with more clusters has a higher probability to be selected as the better one. On the other hand, we have to pay higher cost to search for the relevant documents we really want, since we are facing more clusters now. For an extreme case that FTC returns tens of hundreds of clusters, we cannot be satisfied with this clustering even if the *F*-measure value is near to one. This is one reason why we have introduced the *IMI* measure.

Also, in Fig. 8, the solid line shows the *IMI* values. One observation is that the implication intensity measure can penalize the "ineffective" increase of the cluster number, e.g., from 5 to 15 or from 35 to 50 without any *F*-measure gain. Another observation is that, if the increase of the cluster number can lead to a considerable *F*-measure gain, the *IMI* measure can make a subtle trade-off between the *F*-measure gain and the search cost. For instance, as the cluster number increases from 20 to 25, the *F*-measure gain is dominant, so the implication intensity measure acts similarly to the *F*-measure with a sudden jump. By contrast, however, since the *F*-measure gain is relatively small as the increase of the cluster number from 10 to 20, the implication intensity value drops sharply. Furthermore, comparing the two lines in Fig. 8, it is not difficult to find that the implication intensity measure has a better discrimination than the *F*-measure in the range of $[0, 1]$. To further demonstrate the incremental effect of the *F*-measure, we also present the validation of the *F*-measure and the implication intensity on the `reserve` class of data set `re0`, as shown in Fig. 9. A similar situation holds for this case.

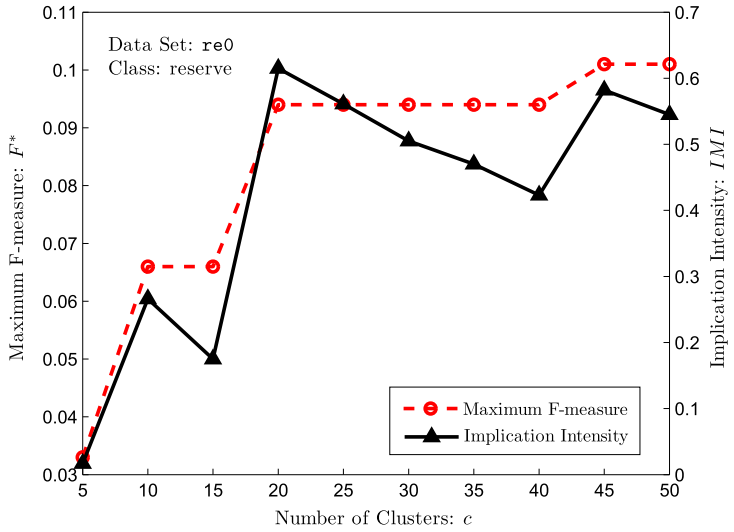

Fig. 9. *IMI* versus maximum *F*-measure: the incremental effect. (Parameters for hyperclique patterns: minimum support – 0.05, minimum *h*-confidence – 0.8).

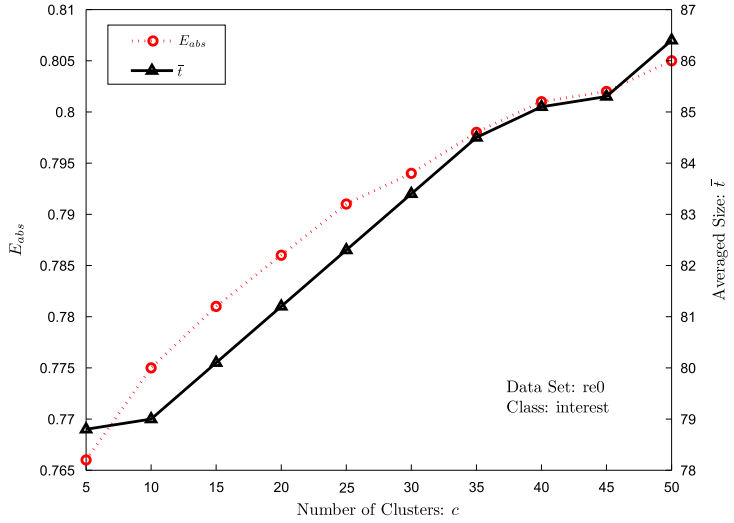

Fig. 10. E_{abs} for the incremental effect.

Table 4
E_{abs} validation for the `reserve` class.

c	5	10	15	20
E_{abs}	89.6	4.24E+05	1.59E+09	3.46E+11

Next, we proceed to investigate the performance of the E_{abs} measure. For the `interest` class of the data set `re0`, E_{abs} increases continuously with the number of clusters *c*, as shown in Fig. 10. This implies that, similar to *IMI*, E_{abs} can also correct the incremental effect. However, one difference is noteworthy; that is, we cannot find any trade-off between the *F*-measure gain and the search cost in E_{abs}. More specifically, unlike *IMI*, the *F*-measure gain seems to have no impact on E_{abs}. One of the reasons for this observation is that the use of $\bar{t}$ may bring some errors into the computations. As can be seen in Fig. 10, $\bar{t}$ is not stable but increases continuously with *c*. This indicates that the cluster sizes are varied to some extent, and the computation of $E(p^*)$ for E_{abs} is therefore questionable.

For the case of the `reserve` class of the data set `re0`, the resultant E_{abs} values are abnormal, as shown in Table 4. When we traced the intermediate results, we found that some variables were overflowed in computing the combinative numbers

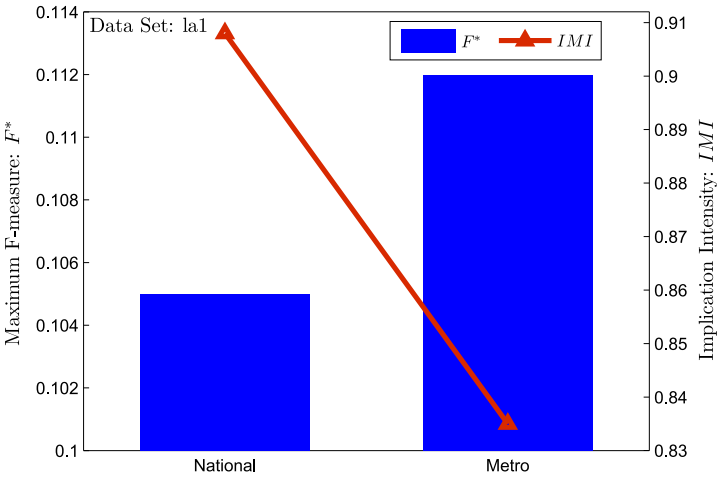

Fig. 11. *IMI* versus maximum *F*-measure: the prior-probability effect. (Parameters for hyperclique patterns: minimum support – 0.05, minimum *h*-confidence – 0.8; FTC parameters: data – `la1`, #cluster – 10).

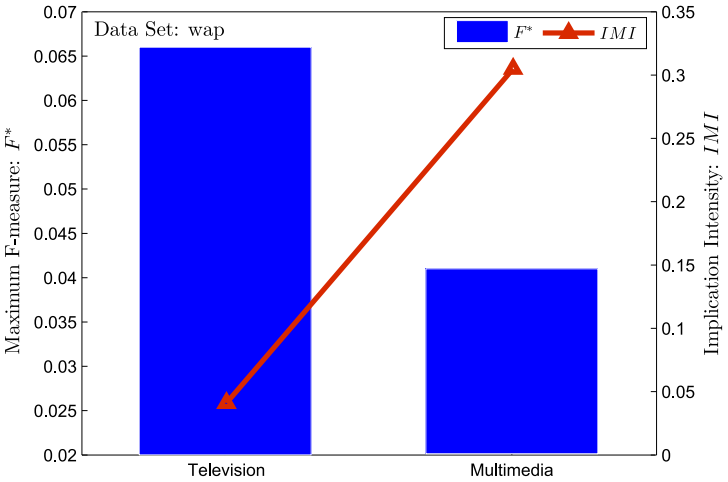

Fig. 12. *IMI* versus maximum *F*-measure: the prior-probability effect. (Parameters for hyperclique patterns: minimum support – 0.05, minimum *h*-confidence – 0.8; FTC parameters: data – `wap`, #cluster – 10).

in Eq. (6). Note that the data type of these variables in our C++ codes is "long double". This case demonstrates that E_{abs} may face difficulties in real-world applications when dealing with extremely large combinative numbers.

In summary, there is the incremental effect if we compare the clustering results with different cluster numbers using the *F*-measure. By contrast, the *IMI* measure has the ability in handling the incremental effect by striking a balance between good and bad effects of the increase of the cluster number.

5.4. The prior-probability effect

Here, we illustrate the prior-probability effect of the *F*-measure and how *IMI* can address this issue.

To show the impact of the incremental effect, we fix the number of clusters as 10 for the data set `la1`, and observe the validation results by the *F*-measure and *IMI*. Fig. 11 shows the comparation results for class `National` and class `Metro`. As can be seen, according to the *F*-measure, the clustering of the `National` documents is worse than the clustering of the `Metro` documents, which contradicts the validation results by *IMI*.

If we take a closer look at the clusters produced, we can find that the number of relevant documents in the optimal cluster of class `National` is only 23, which is smaller than that of class `Metro`: 62. However, the prior-probability of class `Metro` is 0.294, which is much larger than 0.085 - the probability of class `National`. Therefore, it is much easier to find a cluster containing 62 relevant documents of class `Metro` than the search for a cluster containing 23 relevant documents of class `National`. As a result, from a statistical point of view, the clustering for the `Metro` class is indeed worse than the clustering for the `National` class, as indicated by the *IMI* measure in Fig. 11.

To further demonstrate the prior-probability effect of the *F*-measure, we also compare the use of *F*-measure and *IMI* measures on the `Television` and `Multimedia` classes of the data set `wap`, as shown in Fig. 12. Again, the significant disagreement of these two measures indicates the prior-probability effect of the *F*-measure.

In summary, there is the prior-probability effect if we compare the clustering results with different prior probabilities using the *F*-measure. However, the *IMI* measure has the advantage of handling the prior-probability effect by taking into consideration the prior probability of relevant documents.

6. Conclusions

In this paper, we investigated the issues related to overlapping clustering validation in information retrieval. Specifically, we showed that the *F*-measure could lead to biased view on the clustering results due to the incremental and prior-probability effects inherent to *F*-measure. To address these challenges, we designed an implication intensity (*IMI*) measure which shows the probability that the clustering result is better than the matched result by random clustering. Finally, experimental results have shown the effectiveness of *IMI* on validating the clustering performances.

As for the future work, we will investigate the use of *IMI* for validating the clustering results with multiple classes. Also, the random clustering may not be the best baseline for the validation of text clustering, since the information of data attributes is not considered in random clustering. It is possible to design a better baseline from a Bayesian viewpoint. Finally, we will study the applicability of *IMI* to other types of clustering methods. Also, there are many other cluster validation schemes and measures in the literature which may also have the incremental and prior-probability effects. We plan to carry out a systematic study on these measures.

Acknowledgements

The research was partially supported by the National Natural Science Foundation of China (NSFC) (Nos. 70901002, 70890080, 90924020), the MOE Project of Key Research Institute of Humanities and Social Sciences at Universities (No. 07JJD630005), and the National Science Foundation (NSF) via Grant No. CNS 0831186. In addition, the work was also supported by the Research Center for Contemporary Management at Tsinghua University. Finally, we are grateful to the three anonymous referees for their constructive comments on this paper.

Appendix A. The proof of Proposition 1

Proof. We need to prove that

$$\Pr(p^* = j) = \frac{\sum_{i=1}^{c} \binom{c}{i} \left(\binom{P}{j} \binom{T-P}{t-j} \right)^i \left(\sum_{k=0}^{j-1} \binom{P}{k} \binom{T-P}{t-k} \right)^{c-i}}{\binom{T}{t}^c}.$$

For a cluster with exactly j relevant documents, the number of combinations for the cluster members is

$$A(j) = \binom{P}{j} \binom{T-P}{t-j}.$$

Accordingly, for a cluster with the number of relevant documents less than j, the combination number for the cluster members is

$$B(j) = \sum_{k=0}^{j-1} \binom{P}{k} \binom{T-P}{t-k}.$$

Therefore, for all the c clusters, the combination number for the situation that there exactly i clusters having j relevant documents and other clusters having the number of relevant documents less than j is

$$C(i,j) = \binom{c}{i} A(j)^i B(j)^{c-i}.$$

Since $1 \leqslant i \leqslant c$, the combination number for the maximum number of the relevant documents in any of the c clusters is j can be computed as

$$D(j) = \sum_{i=1}^{c} C(i,j).$$

Furthermore, it is trivial to show that the combination number for choosing c clusters, where each of them with t documents, is $\binom{T}{t}^c$. So we finally have

$$\Pr(p^* = j) = D(j) \Big/ \binom{T}{t}^c.$$

The above completes the proof. □

Remark. In some practices, the number of relevant documents P may be smaller than the cluster size t, which results in the error when computing $\binom{P}{j}$ for $A(j)$ or $\binom{P}{k}$ for $B(j)$ in Proposition 1. To avoid this error, we simply let $\binom{P}{j} \equiv 0$ when $P < j$.

Appendix B. The proof for Proposition 2

Proof. Let f_i denote the F-measure value for cluster i generated by $\mathcal{MRC}$, $1 \leqslant i \leqslant c$, we have

$$f^* = \max_{1 \leqslant i \leqslant c} f_i. \tag{B.1}$$

Therefore, according to the definition of OCV and the cluster-independence property of $\mathcal{MRC}$, we have

$$\Pr(f^* < F^*) = \prod_{i=1}^{c} \Pr(f_i < F^*). \tag{B.2}$$

So we turn to the computation of $\Pr(f_i < F^*)$. Consider that

$$f_i = \frac{2p_i}{P + t_i}, \tag{B.3}$$

we have

$$\Pr(f_i < F^*) = \Pr(p_i < 0.5F^*(P + t_i)). \tag{B.4}$$

Let $A_i \equiv \lceil 0.5F^*(P + t_i) \rceil$, then

$$\Pr(f_i < F^*) = \Pr\left(\bigcup_{l=0}^{A_i - 1} \{p_i = l\}\right) = \sum_{l=0}^{A_i - 1} \Pr(p_i = l). \tag{B.5}$$

According to the classic probability model [23]

$$\Pr(p_i = l) = \frac{\binom{P}{l}\binom{T - P}{t_i - l}}{\binom{T}{t_i}}. \tag{B.6}$$

The computation of the combination numbers in Eq. (B.6) will exceed the capacity of the computational system when T is very large. Alternatively, we can approximate the computation by the assumption of the sampling with replacement. This means that p_i is in a binomial distribution with the parameters $n = t_i$ and $p = P/T$, i.e., $p_i \sim B(t_i, P/T)$. Therefore

$$\Pr(p_i = l) \approx \binom{t_i}{l}\left(\frac{P}{T}\right)^l \left(1 - \frac{P}{T}\right)^{t_i - l}. \tag{B.7}$$

If we substitute Eq. (B.7) into Eq. (B.5), we have

$$\Pr(f_i < F^*) \approx \sum_{l=0}^{A_i - 1} \binom{t_i}{l}\left(\frac{P}{T}\right)^l \left(1 - \frac{P}{T}\right)^{t_i - l} = 1 - \sum_{l=A_i}^{t_i} \binom{t_i}{l}\left(\frac{P}{T}\right)^l \left(1 - \frac{P}{T}\right)^{t_i - l}. \tag{B.8}$$

Next, we would like to introduce the so-called "regularized incomplete Beta function" [30,22] to further simplify the computation in Eq. (B.8). In mathematics, the *Euler beta function*, also called the *Euler integral of the first kind* [4], is a special function defined by

$$B(x, y) = \int_0^1 t^{x-1}(1 - t)^{y-1}\, dt, \tag{B.9}$$

for $\mathrm{Re}(x), \mathrm{Re}(y) > 0$. The incomplete beta function is a generalization of the beta function that replaces the definite integral of the beta function with an indefinite integral, as follows:

$$B(x; a, b) = \int_0^x t^{a-1}(1 - t)^{b-1}\, dt. \tag{B.10}$$

Furthermore, the regularized incomplete beta function is defined in terms of the incomplete beta function and the complete beta function. That is

$$I_x(a,b) = \frac{B(x;a,b)}{B(a,b)}. \tag{B.11}$$

Working out the integral by using the integration by parts for integer values a and b, one can finally have

$$I_x(a,b) = \sum_{j=a}^{a+b-1} \frac{(a+b-1)!}{j!(a+b-1-j)!} x^j (1-x)^{a+b-1-j}. \tag{B.12}$$

Now we turn back to the computation of $\Pr(f_i < F^*)$ in Eq. (B.8). Let $x = P/T$, $a_i = A_i$, and $b_i = t_i - A_i + 1$, according to Eq. (B.12), the probability $\Pr(f_i < F^*)$ in Eq. (B.8) can be simplified as

$$\Pr(f_i < F^*) \approx 1 - I_x(a_i, b_i). \tag{B.13}$$

If we substitute Eq. (B.13) into Eq. (B.2), Eq. (7) follows. Thus we complete the proof. □

Remark. The computation of the *IMI* measure makes two approximations: One is to relax the requirement of sampling without replacement to the sampling with replacement, and the other is to make use of the regularized incomplete beta function to compute the partially accumulated binomial distribution. In addition, we should point out that the parameters a_i and b_i in Eq. (7) must be positive integers in the regularized incomplete beta function. Since $a_i = \lceil 0.5F^*(P+t_i) \rceil$, we know that $a_i > 0$ in most cases. As to $b_i = t_i - \lceil 0.5F^*(P+t_i) \rceil + 1$, however, the situation is more complex. Let $c_i \equiv t_i - 0.5F^*(P+t_i) + 1$, it is easy to know $b_i \leqslant c_i$. Therefore, if we let $c_i \leqslant 0$, i.e., $F^* \geqslant 2(t_i+1)/(P+t_i) = 2 - 2(P-1)/(P+t_i)$, we have $b_i \leqslant 0$, which violates the parameter requirement of the incomplete beta function. And this violation happens when F^* is large enough or t_i is small enough, which can happy for real-world applications. To avoid this problem, we can simply let $I_x(a_i, b_i) = 0$ if $b_i \leqslant 0$.

References

[1] R. Agrawal, T. Imielinski, A. Swami, Mining association rules between sets of items in large databases, in: Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, May 1993, pp. 207–216.
[2] R.M. Alidguliyev, Performance evaluation of density-based clustering methods, Information Sciences 179 (20) (2009) 3583–3602.
[3] N.O. Andrews, E.A. Fox, Recent Developments in Document Clustering, Technical Report TR-07-35, Computer Science, Virginia Tech, 2007.
[4] R.A. Askey, R. Roy, Beta function. <http://dlmf.nist.gov/5/12/>, 2008 (accessed 13.12.08).
[5] F. Beil, M. Ester, X. Xu, Frequent term-based text clustering, in: Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2002, pp. 436–442.
[6] D.R. Cutting, D.R. Karger, J.O. Pedersen, J.W. Tukey, Scatter/gather: a cluster-based approach to browsing large document collections, in: Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 1992, pp. 318–329.
[7] M. Halkidi, Y. Batistakis, M. Vazirgiannis, Cluster validity methods: Part I, SIGMOD Record 31 (2) (2002) 40–45.
[8] M. Halkidi, Y. Batistakis, M. Vazirgiannis, Clustering validity checking methods: Part II, SIGMOD Record 31 (3) (2002) 19–27.
[9] E.-H. Han, D. Boley, M. Gini, R. Gross, K. Hastings, G. Karypis, V. Kumar, B. Mobasher, J. Moore, Webace: a web agent for document categorization and exploration, in: Proceedings of the Second International Conference on Autonomous Agents, 1998.
[10] J. Han, J. Pei, Y. Yin, Mining frequent patterns without candidate generation, in: Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, 2000, pp. 1–12.
[11] A.K. Jain, R.C. Dubes, Algorithms for Clustering Data, Prentice Hall, 1988.
[12] N. Jardine, C.J. van Rijsbergen, The use of hierarchical clustering in information retrieval, Information Storage and Retrieval 7 (1971) 217–240.
[13] B. Larsen, C. Aone, Fast and effective text mining using linear-time document clustering, in: Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1999, pp. 16–22.
[14] Chi-Hoon Lee, Osmar R. Zaiane, Ho-Hyun Park, Jiayuan Huang, Russell Greiner, Clustering high dimensional data: a graph-based relaxed optimization approach, Information Sciences 178 (23) (2008) 4501–4511.
[15] D. Lewis, Reuters-21578. <http://www.daviddlewis.com/resources/testcollections/reuters21578/>, 2004 (accessed 29.12.08).
[16] C.D. Manning, P. Raghavan, H. Schütze, Introduction to Information Retrieval, Cambridge University Press, 2008. <http://www-csli.stanford.edu/hinrich/information-retrieval-book.html> (accessed 16.12.08).
[17] M. Mehlitz, C. Bauckhage, S. Albayrak, Normalizing document cluster evaluation results. <http://www.dai-labor.de/fileadmin/files/publications/mehlitz07b.pdf>, 2007 (accessed 12.12.08).
[18] M. Mehlitz, C. Bauckhage, J. Kunegis, S. Albayrak, A new evaluation measure for information retrieval systems, in: Proceedings of the 2007 IEEE International Conference on Systems, Man, and Cybernetics, 2007, pp. 1200–1204.
[19] M. Meila, Comparing clusterings – an axiomatic view, in: Proceedings of the 22nd International Conference on Machine Learning, 2005, pp. 577–584.
[20] Kweku-Muata Osei-Bryson, Towards supporting expert evaluation of clustering results using a data mining process model, Information Sciences 180 (3) (2010) 414–431.
[21] M.F. Porter, An algorithm for suffix stripping, Program 14 (3) (1980).
[22] W.H. Press, B.P. Flannery, S.A. Teukolsky, W.T. Vetterling, Numerical Recipes in C: The Art of Scientific Computing, second ed., Cambridge University Press, Cambridge, UK, 1992.
[23] S.M. Ross, Introduction To Probability Models, ninth ed., Elsevier India Private Limited, 2007.
[24] M. Steinbach, G. Karypis, V. Kumar, A comparison of document clustering techniques, in: Workshop on Text Mining, The Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2000.
[25] P.-N. Tan, M. Steinbach, V. Kumar, Introduction to Data Mining, Addison-Wesley, 2005.
[26] A. Tombros, R. Villa, C.J. van Rijsbergen, The effectiveness of query-specific hierarchic clustering in information retrieval, Information Processing and Management 38 (4) (2002) 559–582.
[27] TREC, Text retrieval conference. <http://trec.nist.gov/>, 2000 (accessed 16.12.08).
[28] C.J. van Rijsbergen, Information Retrieval, second ed., Butterworths, London, 1979.
[29] Xiaojun Wan, A novel document similarity measure based on earth movers distance, Information Sciences 177 (18) (2007) 3718–3730.

[30] Wikipedia, Beta function – wikipedia, the free encyclopedia. <http://en.wikipedia.org/wiki/Beta_function>, 2008 (accessed 13.12.08).
[31] Wikipedia, Information retrieval – wikipedia, the free encyclopedia. <http://en.wikipedia.org/w/index.php?title=Information_retrieval&oldid=2%50468748>, 2008 (accessed 12.12.08).
[32] J. Wu, H. Xiong, J. Chen, Adapting the right measures for *k*-means clustering, in: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2009, pp. 877–886.
[33] J. Wu, H. Xiong, J. Chen, External validation measures for *k*-means clustering: a data distribution perspective, Expert Systems with Applications 36 (3) (2009) 6050–6061.
[34] J. Wu, H. Xiong, J. Chen, Towards understanding hierarchical clustering: a data distribution perspective, Neurocomputing 72 (10–12) (2009) 2319–2330.
[35] W. Wu, H. Xiong, S. Shekhar, Clustering and Information Retrieval, Kluwer Academic Publishers, 2003.
[36] H. Xiong, P.-N. Tan, V. Kumar, Hyperclique pattern discovery, Data Mining and Knowledge Discovery 13 (2) (2006) 219–242.
[37] H. Xiong, J. Wu, J. Chen, *K*-means clustering versus validation measures: a data-distribution perspective, IEEE Transactions on Systems, Man, and Cybernetics: Part B 39 (2) (2009) 318–331.
[38] Hui Xiong, Pang-Ning Tan, Vipin Kumar, Mining strong affinity association patterns in data sets with skewed support distribution, in: Proceedings of the Third IEEE International Conference on Data Mining (ICDM 2003), December 19–22, IEEE Computer Society, Melbourne, Florida, USA, 2003, pp. 387–394.
[39] Yahoo!, Yahoo! home. <http://www.yahoo.com>, 2008 (accessed 29.12.08).
[40] O. Zamir, O. Etzioni, Web document clustering: a feasibility demonstration, in: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 1998, pp. 46–54.
[41] Y. Zhao, G. Karypis, Criterion functions for document clustering: experiments and analysis, Machine Learning 55 (3) (2004) 311–331.
[42] H.-T. Zheng, B.-Y. Kang, H.-G. Kim, Exploiting noun phrases and semantic relationships for text document clustering, Information Sciences 179 (13) (2009) 2249–2262.