Information-theoretic Term Weighting Schemes for Document Clustering

Weimao Ke
Lab for Information Network & Computing Studies
College of Information Science and Technology
Drexel University, Philadelphia, PA 19104, U.S.A.
wk@drexel.edu

ABSTRACT

We propose a new theory to quantify information in probability distributions and derive a new document representation model for text clustering. By extending Shannon entropy to accommodate a non-linear relation between information and uncertainty, the proposed Least Information theory (LIT) provides insight into how terms can be weighted based on their probability distributions in documents vs. in the collection. We derive two basic quantities in the document clustering context: 1) LI Binary (LIB) which quantifies information due to the observation of a term's (binary) occurrence in a document; and 2) LI Frequency (LIF) which measures information for the observation of a randomly picked term from the document. Both quantities are computed given term distributions in the document collection as prior knowledge and can be used separately or combined to represent documents for text clustering. Experiments on four benchmark text collections demonstrate strong performances of the proposed methods compared to classic TF*IDF. Particularly, the LIB*LIF weighting scheme, which combines LIB and LIF, consistently outperforms TF*IDF in terms of multiple evaluation metrics. The least information measure has a potentially broad range of applications beyond text clustering.

Categories and Subject Descriptors

H.3.1 [**Information storage and retrieval**]: Content Analysis and Indexing; H.3.3 [**Information storage and retrieval**]: Information Search and Retrieval—*Clustering*

General Terms

Theory, Algorithms, Performance, Experimentation

Keywords

term weighting, information measure, semantic information, document representation, text clustering

JCDL'13, July 22–26, 2013, Indianapolis, Indiana, USA.

1. INTRODUCTION

Clustering, or unsupervised classification, is the process of bringing like entities together [12]. Text clustering is focused on partitioning unstructured text documents and underlies many applications for information organization and knowledge management [6]. Document clustering also supports important processes in text mining and information retrieval [21, 33].

In text clustering research, TF*IDF has been extensively used for term weighting and document representation [21, 34]. While term frequency (TF) indicates the degree of a document's association with a term, inverse document frequency (IDF) is the manifestation of a term's specificity, key to determine the term's value toward weighting and relevance ranking [15]. While many clustering algorithms have been developed, TF*IDF and its variations remain the *de facto* standard for term weighting in text clustering [33, 21, 34, 14]. Given strong empirical performances of classic TF*IDF, limited research has focused on innovation of term weighting schemes.

Information and probability theories have provided important guidance on the development of classic techniques such as probabilistic and language modeling [27]. Information-theoretic measures such as mutual information and Kullback-Leibler (KL) divergence[1] have also been used for various processes including feature selection and matching [18, 33].

The probabilistic retrieval framework provides an important theoretical ground to IDF weights [26]. IDF in the form of $-\ln \frac{n_i}{N}$, where n_i is the number of documents containing term i among N total number of documents, resembles the entropy formula in Shannon's information theory. Several works have attempted to justify IDF from an information-theoretic view. IDF can be viewed as Kullback-Leibler (KL) information (*relative entropy*) between term probability distributions in a document and in the collection [1].

From an information-centric view, this research aims to develop a new model for term weighting and document representation. By quantifying the amount of information required to explain probability distribution changes, the proposed *least information* theory (LIT) establishes a new basic information quantity and provides insight into how terms can be weighted based on their probability distributions in documents vs. in the collection. We derive two basic quanti-

[1]The literature has used a variety of names in reference to *KL divergence*. While Kullback preferred *discrimination information* for the principle of minimum discrimination information (MDI) [17], the literature has often referred to it as divergence information or relative entropy.

ties, namely LI Binary (LIB) and LI Frequency (LIF), which can be used separately or combined to represent documents. We conduct experiments on several benchmark collections for text clustering to demonstrate the proposed methods' effectiveness compared to TF*IDF. The major contribution here is more than another term weighting scheme that is empirically competitive. More important is the new *least information* (LIT) measure that can be used to attack many other problems related to quantifying semantic amounts of information.

2. PROPOSED THEORY

In this section, we propose a new theory to quantify meaning of information via extension of Shannon's entropy equation. We start with a discussions on issues of existing information measures, what to expect about the desired information quantity, and introduce the *least information* theory (LIT) in which expected characteristics are observed.

2.1 Information Measures and Problems

Shannon entropy measures uncertainty as a property of a probability distribution whereas the amount of (missing) information is a function of linear uncertainty reduction [29]. The underlying assumption for this entropy-information relation is that information (always) reduces uncertainty. The amount of information is determined by a specified probability distribution regardless of the ultimate outcome [5].

In reality, however, there are many situations in which information does not necessarily reduce uncertainty – uncertainty may increase or decrease due to new information. In addition, the amount of information depends not only on the overall uncertainty change but also on how individual probabilities vary. For example, an unlikely event being the ultimate outcome requires more explanation (information) than in the case of a very likely event happening.

Different amounts of information are needed to explain different (and perhaps opposite) outcomes. We reason that, while uncertainty is a property of a specified probability distribution, the amount of information required to explain an outcome and more generally to explain a change in the probability distribution is interpretation/meaning-dependent and is more complex than a linear function of uncertainty.

Indeed, using Shannon's entropy measure to quantify the amount of *meaningful* information (with proper interpretation) is beyond the scope of classic information theory. The original purpose of Shannon's theory was for engineering communication systems where the "*meaning* of information was considered irrelevant" [29, p. 379]. As Rapoport (1953) put it, it is about technical problems that can be treated independently of the semantic content of messages [25].

Digital libraries technologies such as those related to information organization and retrieval deal with issues of semantics and relevance, beyond pure engineering problems. Measuring *semantic quantities* of information requires innovation on the theory, better clarification of the relationship between information and entropy, and justification of this relationship.

While related quantities such as KL information (*relative entropy*) offer alternatives to the simplified entropy reduction view of information, some characteristics of *relative entropy* do not meet our expectations about such a measure. Specifically, the asymmetry of the KL function due to the assumption about a benchmark distribution in the evaluation disqualifies it as a metric [8].

In addition, *relative entropies* over the course of continuous probability changes in one direction do not add up to the overall amount. Finally and very important, extreme probability changes (e.g., when an event changes from being very unlikely to almost certainty) lead to infinite KL information, which is a particularly undesirable property for term weighting. We address these issues in the proprosed *least information theory* (LIT) below.

2.2 Least Information Theory (LIT)

In this section, we present the *least information* theory (LIT) to quantify meaning (semantics) in probability distribution changes. Here we shall clarify on the definition of *meaning* by reusing D. M. MacKay's terms, in which the meaning of information is defined as a "selective function on a range of the recipient's states of conditional readiness for goal-directed activity." [23, p.24] Meaning is a relationship between information and the recipient rather than a property alone. Information is not restricted to the amount in message transmission but is subject to interpretation [11]. Introduction of information results in changes of readiness states or belief (as in probabilities). In this view, meaning can (at least in part) be quantified by what varies in the beliefs or estimated probabilities of inferences.

Let X be prior (initially specified) probabilities for a set of exhaustive and mutually exclusive inferences: $X = [x_1, x_2, .., x_n]$, where x_i is the prior probability of the i^{th} inference on a given hypothesis. Let Y denote posterior (changed) probabilities after certain information is known: $Y = [y_1, y_2, .., y_n]$, where y_i is the *informed* probability of the i^{th} inference. Uncertainties/entropies of the two distributions can be computed by Shannon entropy:

$$H(X) = -k\sum_{i=1}^{n} x_i \ln x_i \quad (1)$$

$$H(Y) = -k\sum_{i=1}^{n} y_i \ln y_i \quad (2)$$

The amount of information obtained from X to Y, in Shannon's treatment, can be measured via the reduction of entropy:

$$\Delta H = H(X) - H(Y) \quad (3)$$

The inferences are exclusive and involve different meanings. When probabilities X to Y are not identical, the two distributions are semantically different and it is obvious that some amount of information is responsible for the variance. We can examine the amount of information associated with individual inferences via the measurement of uncertainty change. With Equation 3, however, it is easy to show that when there are changes in the probabilities, there may be increases, decreases, or no change in the overall uncertainty. We observe that even when there is no change in the entropy, there is still an amount of information responsible for any variance in the probability distribution. To use the overall (system-wide) uncertainty for the measurement of information ignores semantic relevance of changes in individual inferences.

Here our new *least information* model departs from the classic measure of information as reduction of uncertainty (entropy). First, we reason that any change in the uncertainty of an inference, either an increase or decrease, requires a relevant amount of information that is responsible for it. The overall information needed to explain changes in all inference probabilities is the sum of individual pieces of information associated with each inference.

Second, for an individual inference i, the probability may vary in one of the two semantic directions, i.e., to increase or to decrease the likelihood. In either case, there is always a (positive) amount of information responsible for that variance. If we assume inferences are semantically independent[2], the absolute values of these independent pieces of information add linearly to the overall amount of information.

In addition, it is reasonable for such an information quantity to meet the condition that continuous, smaller changes in one direction add incrementally to a bigger change in the same direction. That is, pieces of information responsible for small, continuous changes of an inference probability in the same direction should add up to the amount of information for the overall change. For example, if the i^{th} inference's probability increases from x_i to y_i and then to z_i, the (least) amount of information required for the change from x_i to y_i and the amount from y_i to z_i should add up to the overall (least) information required for the change from x_i to z_i.

We define dH_i as the amount of entropy change due to a tiny change dp_i of probability p_i:

$$dH_i \quad = \quad -\ln p_i dp_i \tag{4}$$

In the configuration view of entropy, this microscopic variance of entropy due to a small change in an inference's probability is the change of the weighted (p_i) number of configurations ($\ln \frac{1}{p_i}$) [10, 5]. In other words, it is the change in the number of configurations ($\ln \frac{1}{p_i}$) due to a varied probability weight (p_i).

Every tiny change in the probabilities requires some explanation (information). Aggregating (integrating) the small changes of uncertainty leads to the amount of information required for a macro-level change. A macroscopic uncertainty change due to a significant probability shift of an inference is the sum (integration) of continuous microscopic changes in the variance range. Therefore, we define the least amount of information I_i required to explain the probability change of the i^{th} inference as the integration (aggregation) of all tiny absolute (positive) changes of entropy dH_i:

$$\begin{aligned} I_i &= \Big| \int_{x_i}^{y_i} dH_i \Big| \\ &= \Big| \int_{x_i}^{y_i} -\ln p_i dp_i \Big| \\ &= \Big| p_i(1-\ln p_i) \Big|_{x_i}^{y_i} \qquad (5) \\ &= \Big| y_i(1-\ln y_i) - x_i(1-\ln x_i) \Big| \qquad (6) \end{aligned}$$

[2]Inference probabilities are never perfectly independent of one another given the degree of freedom. But to simplify the discussion and formulation, we use the independence assumption.

where x_i is the initial probability of the i^{th} inference and y_i the posterior probability of the same inference. We define *informative entropy* g_i as a function of an inference's probability:

$$g_i \quad = \quad p_i(1-\ln p_i) \tag{7}$$

The equation for *least information* I_i for the i^{th} inference in Equation 6 can be rewritten as:

$$I_i \quad = \quad \Big| g(y_i) - g(x_i) \Big| \tag{8}$$

The total *Least Information* I is the sum of partial least information for every inference:

$$\begin{aligned} I &= \sum_{i=1}^{n} I_i \\ &= \sum_{i=1}^{n} \Big| g(y_i) - g(x_i) \Big| \\ &= \sum_{i=1}^{n} \Big| y_i(1-\ln y_i) - x_i(1-\ln x_i) \Big| \qquad (9) \end{aligned}$$

where n is the number of inferences, x_i is the initially specified probability of the i^{th} inference, and y_i the revised probability of the i^{th} inference.

2.3 Important Model Characteristics

It is worth noting that Equation 9 is to measure the *least amount* of information required to explain a probability distribution change for a set of inferences. Here is why we include the word *least* in the nomenclature. Given that information may alter a probability distribution in various semantic directions and change the uncertainty in both positive and negative directions, the actual amount of information leading to such a change may consist of multiple pieces of information acting in different directions.

Without an exhaustive analysis of the process, the actual amount of information cannot be deduced solely from an investigation of probability distributions. It is only reasonable to quantify the *least information* needed for that change – that is, the sum of all needed amounts of information at the very least, every tiny piece of which contributes in the same direction of a change. In addition, this model does not consider the process of removing information, which, in effect, is equivalent to adding another piece of information that has perfectly opposite semantics[3] in the same amount.

Based on Equation 9, several important characteristics of *least information* can be observed. Figure 1 compares the *least information* measure with entropy reduction and relative entropy in a two-exclusive-inference case. We summarize some of these characteristics below.

- Absolute information and symmetry: The amount of *least information* required for a probability change from X to Y is the same as that from Y to X, though their semantic meanings are different.

[3]The term *opposite* does not indicate true vs. false information. Opposite information semantics are essentially to increase vs. to decrease the probability of an inference, e.g., good news vs. bad news about a candidate that may influence the outcome of an election.

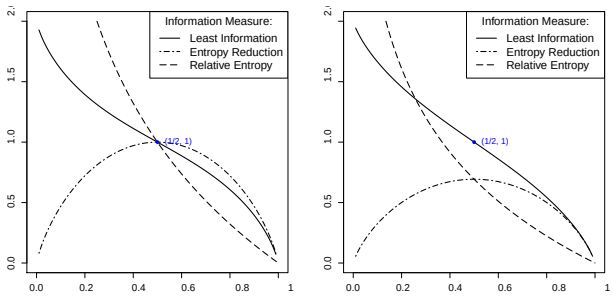

Figure 1: Least Information vs. Entropy: Reducing two exclusive uncertain inferences to certainty. X denotes P_1 which is the probability for 1st inference of 2 mutually exclusive. Y denotes the amount of information (I) in terms of each measure. The 1st inference is assumed to be the ultimate outcome in the figure. The asymmetry of *least information* in the plot is a manifestation of its dependence on the outcome. Compare to Fig. 7 in Shannon (1948).

- Addition of continuous change: Amounts of *least information* for small, continuous probability changes in the same semantic directions add linearly to the amount of *least information* responsible for the overall change. In short, $I(X \rightarrow Z) = I(X \rightarrow Y) + I(Y \rightarrow Z)$, if and only if $X \rightarrow Y$ and $Y \rightarrow Z$ are in the same semantic direction.
- Unit Information: In the special case when there are two equally possible inferences, the amount of *least information* needed to explain an outcome (certainty) is one: $I(p_1 = p_2 = \frac{1}{2} \rightarrow p_1 = 1) = 1$, regardless of the log base in the equation (see data point $(\frac{1}{2}, 1)$ in Figure 1).
- In the special case of reducing uncertain inferences to certainty (with the ultimate case):
  - With equally likely inferences, when there are more choices, the least information needed to explain an outcome is larger.
  - The less likely the outcome, the larger the amount of *least information* needed to explain it.
- Zero least information: The amount of *least information* is zero if and only if there is no change in the probability distribution (identical distributions).

2.4 Least Information for Term Weighting

Now we apply the proposed *least information* theory to term weighting and document representation. A text document can be viewed as a set of terms with probabilities (estimated by frequencies) of occurrence. We conjecture that the larger amount *least information* is needed to explain a term's probability in a document (vs. in the collection), the more heavily the term should be weighted to represent the document [2]. Hence, we transform the question of document representation to weighting terms according to their amounts of *least information* in documents.

In this study, we propose two specific weighting methods, one based on a binary representation of term occurrence (0 vs. 1) and the other based on term frequencies. These two methods will be used separately and combined in fusion methods as well. We use maximum likelihood estimates (MLE) to get term probabilities based on observed data [24].

2.4.1 LI Binary (LIB) Model

In the binary model, a term either occurs or does not occur in a document. If we randomly pick a document from the collection, the chance that a term t_i appears in the document can be estimated by the ratio between the number of documents containing the term n_i (i.e., document frequency) and the total number of documents N using MLE. Let $p(t_i|C) = n_i/N$ denote the probability of term t_i occurring in a randomly picked document in collection C; $p(\bar{t_i}|C)$ is the probability that the term does not appear:

$$p(\bar{t_i}|C) = 1 - p(t_i|C) = 1 - n_i/N$$

When a specific document d is observed, it becomes certain whether a term occurs in the document or not. Hence the term probability given a specific document $p(t_i|d)$ is either 1 or 0. Given the definition of g_i in Equation 7, the least amount of information in term t_i from observing document d can be computed by:

$$\begin{aligned} I(t_i, d) &= \Big|g(t_i|d) - g(t_i|C)\Big| \\ &\quad + \Big|g(\bar{t_i}|d) - g(\bar{t_i}|C)\Big| \end{aligned} \tag{10}$$

The above equation gives the amount of information a term conveys in a document regardless of its semantic direction. When a query term t_i does not appear in document d, the least information associated with the term should be treated as *negative* because it makes the document less relevant to the term. Hence, the ranking function should not only consider the amount of information but also the *sign* (positive vs. negative) of the quantity. Hence, LI Binary (LIB) can be computed by:

$$\begin{aligned} LIB_2(t_i, d) &= g(t_i|d) - g(t_i|C) \\ &\quad - g(\bar{t_i}|d) + g(\bar{t_i}|C) \end{aligned} \tag{11}$$

For term weighting, we are more interested in the likelihood of a term appearing in a document. Keeping only quantities related to t_i (and removing those associated with $\bar{t_i}$), we simplify the LIB equation to:

$$LIB(t_i, d) = g(t_i|d) - g(t_i|C) \tag{12}$$

$$= g(t_i|d) - \frac{n_i}{N}\Big(1 - \ln\frac{n_i}{N}\Big) \tag{13}$$

The quantity depends on the observation of term t_i in the document: $g(t_i|d)$ is 1 when t_i appears in document d and 0 if otherwise, according to Equation 7. That is:

$$LIB(t_i, d) = \begin{cases} 1 - \frac{n_i}{N}\Big(1 - \ln\frac{n_i}{N}\Big) & t_i \in d \\ -\frac{n_i}{N}\Big(1 - \ln\frac{n_i}{N}\Big) & t_i \notin d \end{cases} \tag{14}$$

where n_i is the document frequency of term t_i and N is the total number of documents. The larger the LIB, the more information the term contributes to the document and

should be weighted more heavily in the document representation. LIB is similar in spirit to IDF and its value represents the discriminative power of the term when it appears in a document.

2.4.2 LI Frequency (LIF) Model

In the LI Frequency (LIF) model, we use term frequencies to model *least information*. Treating a document collection C as a meta-document, the probability of a term randomly picked from the collection being a specific term t_i can be estimated by: $p(t_i|C) = F_i/L$, where F_i is the total number of occurrences of term t_i in collection C and L the overall length of C (i.e., the sum of all document lengths).

When a specific document d is observed, the probability of picking term t_i from this document can be estimated by: $p(t_i|d) = tf_{i,d}/L_d$, where $tf_{i,d}$ is the number of times term t_i occurs in document d and L_d is the length of the document. Again, for each term t_i, there are two exclusive inferences, namely the randomly picked term being the specific term (t_i) or not ($\bar{t_i}$). To quantify a term's LIF weight, we measure *least information* that explains the change from the term's probability distribution in the collection to its distribution in the document in question:

$$\begin{aligned} LIF_2(t_i, d) &= g(t_i|d) - g(t_i|C) \\ &\quad + g(\bar{t_i}|C) - g(\bar{t_i}|d) \end{aligned} \tag{15}$$

We focus on the quantities $g(t_i|d)$ and $g(t_i|C)$ to estimate *least information* of each term when a specific document is observed. Without quantities $g(\bar{t_i}|C)$ and $g(\bar{t_i}|d)$, the LIF equation is simplified to:

$$\begin{aligned} LIF(t_i, d) &= g(t_i|d) - g(t_i|C) \qquad (16) \\ &= \frac{tf_{i,d}}{L_d}(1 - \ln \frac{tf_{i,d}}{L_d}) \\ &\quad - \frac{F_i}{L}(1 - \ln \frac{F_i}{L}) \qquad (17) \end{aligned}$$

where $tf_{i,d}$ is term frequency of term t_i in document d and L_d is the document length. F_i is collection frequency of term t_i (the sum of term frequencies in all documents) whereas L is the overall length of all documents. In a sense, LIF can be seen as a new approach to modeling term frequencies with document length and collection frequency normalization. In this study, we use raw term frequencies with MLE to estimate probabilities and do not use any smoothing techniques to fine tune the estimates.

2.4.3 Fusion of LIB & LIF

While LIB uses binary term occurrence to estimate least information a document carries in the term, LIF measures the amount of least information based on term frequency. The two are related quantities with different focuses. As discussed, the LIB quantity is similar in spirit to IDF (inverse document frequency) whereas LIF can be seen as a means to normalize TF (term frequency).

In light of TF*IDF, we reason that combining the two will potentiate each quantity's strength for term weighting. Hence we propose three fusion methods to combine the two quantities by addition and multiplication:

1. LIB+LIF: To weight a term, we simply add LIB and LIF together by treating them as two separate pieces of information.
2. LIB*LIF: In this fusion method, we follow the idea of TF*IDF by multiplying LIB and LIF quantities for each term. Because a *least information* quantity falls in the range of $[-1, 1]$ and can be a negative value, we normalize LIB and LIF values to $[0, 2]$ by adding 1 to each before multiplication.
3. LIB*TF: This method multiplies the LIB quantity by a document length normalized TF (term frequency), similar to the above LIB*LIF method.

These fusion methods allow us to examine potential strengths and weaknesses of the proposed *least information* term weights for clustering. We study LIB and LIF as well as the above fusion methods in experiments. And given the extensive use of TF*IDF in text clustering research, we use it for comparison in the study.

3. EXPERIMENTAL SETUP

3.1 Data Collections

Several benchmark collections were used in the study to evaluate the effectiveness of proposed term weighting methods for text clustering. Some of these collections, including the WebKB 4 universities data, the 20 Newsgroups collection, and the RCV1 Reuters corpus, had been widely used for text clustering and classification research. The New York Times annotated corpus was a relatively new development and had not been extensively adopted for clustering experiments.

- WebKB 4 Universities Data (WebKB): This data set contains 8,282 web pages collected in 1997 from computer science departments of various universities, which were manually categorized into seven categories such as student, faculty, and department. This was developed by the WebKB project at CMU [9].
- 20 Newsgroups (20News): The collection contains 20,000 messages from 20 news groups (categories) [19]. The messages were randomly picked to distribute evenly among the categories. We used a revised version which retained 18,828 messages after duplicate removal. We used all 20 categories as gold standard labels.
- Reuters Corpus Volume 1 (RCV1-v2): The RCV1 collection contains 804,414 newswire stories made available by Reuters. RCV1-v2 is a corrected version of the original collection, in which documents were manually assigned to a hierarchy of 103 categories [20]. There are four top-level categories (under the hierarchical root), which we used as labels for evaluation.
- New York Times Annotated Corpus (NYTimes): The NYTimes corpus contains more than 1.8 million articles in New York Times from 1987 - 2007 [28]. The corpus is very rich in human annotation such as summaries and subject descriptors. For each year, we identified articles assigned to one and only one *taxonomic classifier* and used the third-level categories under *top/news/* (e.g., *top/news/science* and *top/news/business*)

as labels. After several categories such as *corrections* were removed, the final data set contained 179,175 documents in 15 categories.

3.2 System Settings

We developed an experimental clustering system based on the Weka data mining framework [32]. We implemented various document representation methods including the proposed term weighting schemes and TF*IDF based on a Weka vectorization filter. Existing implementations in the framework for k-means clustering [3] and hierarchical agglomerative clustering (HAC) were reused in experiments [35]. We tokenized documents into single words, removed stop-words, and normalized terms using an iterated Lovins stemmer [22]. A number of most frequent words were selected as features (DF thresholding); 1,000 features were used in main experiments. We varied the number of features in experiments to study the influence of feature selection. All documents were normalized to unit vectors.

In k-means clustering, we used the euclidean distance and set the maximum number of iterations to 200. We conducted 30 runs of k-means for each experimental setting, in which clustering was performed on a random sample of 2,000 documents. The HAC clustering used *complete link* and angle distance (based on an arc cosine function). For each experimental setting, HAC was performed on a random sample of 1,000 documents for 20 runs. We set the number of desired clusters to the number of classes/labels in each data collection.

3.3 Evaluation Metrics

Using categorical labels available in data as the gold standard, we evaluated clustering results based on several classic metrics, namely, *purity*, *rand index*, *precision*, *recall*, and F_1. By assigning each cluster to the most frequent class (label) in it, we can compute *purity* by:

$$purity(C, L) = \frac{1}{N}\sum_i \max_j |c_i \cap l_j| \quad (18)$$

where N is the total number of documents. C is the set of clusters and c_i is the i^{th} cluster. L is the set of labels (classes) where l_j is the j^{th} label.

Computing the other metrics such as *rand index* is by viewing document clustering as a series of decision making [24]. Given the following table which summarizes the numbers of correctly and incorrectly clustered document pairs:

Table 1: Decision table of clustering

System⇒ Labels⇓	**Same Cluster**	**Diff Clusters**
Same Class	*TP*: True Positive	*FN*: False Negative
Diff Classes	*FP*: False Positive	*TN*: True Negative

Rand Index measures the ratio of correct decisions and is computed by:

$$RI = \frac{TP + TN}{TP + FP + FN + TN} \quad (19)$$

Likewise, precision, recall, and F_1 can by computed by:

$$P = TP/(TP + FP) \quad (20)$$

$$R = TP/(TP + FN) \quad (21)$$

$$F_1 = 2 * P * R/(P + R) \quad (22)$$

Whereas *rand index* measures clustering accuracy by taking into account both true positive and true negative, classic IR evaluation metrics such as precision and recall emphasize the ability to find relevant answers/pairs (true positive). Purity and precision are similar in that they both focus on the internal accuracy within each cluster. Recall, on the other hand, addresses the effectiveness of having as many relevant document pairs as possible in one cluster. With these various metrics, we were able to examine strengths and weaknesses of the proposed methods in multiple perspectives.

4. RESULTS

In each set of experiments presented here, best scores in each metric are highlighted in **bold** whereas *italic* values are those better than TF*IDF baseline scores. A significant better result according to t-test at 0.05 is shown with an asterisk (∗). We first present results using k-means clustering with various weighting schemes and with 1,000 features in sections 4.1 - 4.4. Analysis of the impact of feature selection on clustering effectiveness in section 4.5 will show overall best results were obtained with 1,000 features in WebKB, 20News, and RCV1 collections. We discuss hierarchical agglomerative clustering (HAC) results in section 4.6.

4.1 WebKB 4 Universities Data

Table 2 shows k-means clustering results on the WebKB 4 Universities data set. The proposed methods LIB, LIB+LIF, and LIB*LIF all outperformed TF*IDF in terms of purity, rand index, and precision. LIF and LIB*TF, which have an emphasis on term frequency, achieved significantly better recall scores.

Table 2: WebKB with k-means clustering

Method	Purity	RIndex	Prec.	Recall	F_1
TF*IDF	0.504	0.655	0.343	0.245	0.283
LIB	*0.520*∗	**0.686**∗	**0.383**∗	0.215	0.275
LIF	0.455	0.600	0.282	*0.277*∗	0.276
LIB*TF	0.459	0.571	0.273	**0.323**∗	**0.292**
LIB+LIF	*0.517*∗	*0.680*∗	*0.377*∗	0.214	0.272
LIB*LIF	**0.524**∗	*0.681*∗	*0.376*∗	0.208	0.268

4.2 20 Newsgroups Data

K-means clustering with the 20 Newsgroups data, as shown in Table 3, presents a slightly different picture. While best scores were achieved among the proposed methods, LIF appeared to have produced most of the best scores, with purity, precision, and F_1 scores significantly higher than those of TF*IDF. Overall, k-means was not very effective in terms of within-cluster accuracy (purity and precision).

4.3 RCV1 Reuters Corpus

K-means clustering experiments on the RCV1 corpus showed a pattern quite similar to WebKB results. As shown in

Table 3: 20 Newsgroup with k-means clustering

Method	Purity	RIndex	Prec.	Recall	F_1
TF*IDF	0.229	0.771	0.0973	0.350	0.145
LIB	*0.241*	**0.832**	*0.111*∗	0.286	*0.154*
LIF	**0.269**∗	*0.816*	**0.117**∗	*0.372*	**0.173**∗
LIB*TF	0.174	0.631	0.0757	**0.489**∗	0.126
LIB+LIF	*0.231*	0.732	*0.106*	*0.408*	*0.155*
LIB*LIF	0.214	*0.793*	0.0957	0.309	0.137

Table 4, the proposed methods outperformed TF*IDF in terms of multiple metrics. Compared to TF*IDF, LIB*LIF, LIB+LIF, and LIB performed significantly better in purity, rand index, and precision whereas LIF and LIB*TF achieved significantly better scores in recall.

Table 4: RCV1 with k-means clustering

Method	Purity	RIndex	Prec.	Recall	F_1
TF*IDF	0.699	0.678	0.511	0.576	0.531
LIB	*0.733*∗	*0.711*∗	*0.543*	0.538	*0.539*
LIF	*0.719*	*0.699*	*0.546*	*0.663*∗	**0.588**∗
LIB*TF	0.650	0.637	0.468	**0.670**∗	*0.542*
LIB+LIF	*0.725*	*0.711*∗	*0.544*	0.554	*0.547*
LIB*LIF	**0.741**∗	**0.713**∗	**0.549**∗	0.531	*0.539*

4.4 New York Times Collection

With the NY Times corpus, LIB*LIF continued to dominate best scores and performed significantly better than TF*IDF in terms of purity, rand index, and precision (Table 5). This is very consistent with WebKB and RCV1 results. While LIB and LIB+LIF did well in terms of rand index, LIF and LIB*TF were competitive in recall.

Table 5: NYTimes with k-means Clustering

Method	Purity	RIndex	Prec.	Recall	F_1
TF*IDF	0.825	0.632	0.683	0.383	0.456
LIB	0.819	*0.634*	*0.691*	*0.399*∗	*0.469*
LIF	0.807	0.627	0.662	*0.433*∗	**0.481**∗
LIB*TF	0.813	0.622	0.672	**0.435**∗	*0.473*∗
LIB+LIF	*0.829*	*0.640*	*0.700*	*0.405*∗	*0.476*∗
LIB*LIF	**0.837**∗	**0.647**∗	**0.719**∗	*0.391*	*0.472*∗

Overall, LIB*LIF had a strong performance across the data collections. Methods with the LIB quantity, especially LIB, LIB+LIF, and LIB*LIF, were effective when the evaluation emphasis was on within-cluster (internal) accuracy, e.g., in terms of purity and precision. Similar to IDF, LIB was designed to weight terms according to their *discriminative* powers or *specificity* in terms of Sparck Jones [15]. Hence, it helped improve precision-oriented effectiveness. The other methods such as LIF and LIB*TF emphasize *term frequency* in each document and, with the ability to associate one document to another by assigning term weights in a *less discriminative* manner, were able to achieve better recalls.

4.5 Impact of Feature Selection

Now we look at the impact of feature selection on the effectiveness of document representation for clustering. In this work, we selected features based on their frequencies in the collection (DF thresholding), which was computationally simple and had been found in several studies to be a very effective feature selection technique for clustering and categorization [21, 33, 34]. We varied the number of features N_f in each set of experiments, for which the top N_f most frequent terms were kept for document representation. We performed this experimental analysis on three collections, namely, WebKB, 20Newsgroup, and RCV1.

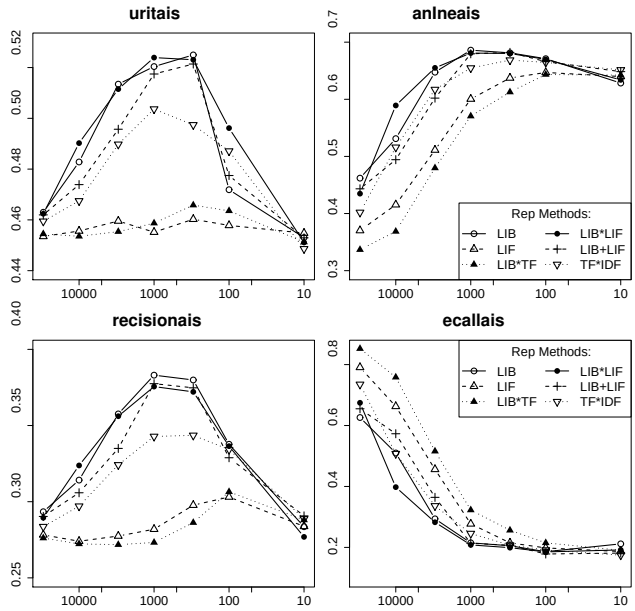

Figure 2: WebKB: Impact of feature selection. X denotes # of features and is log-transformed. Y is the metric score.

Figure 2 shows the influence of the number of selected features on clustering effectiveness with the WebKB data. Note that the X axis is logarithmic and the number of features N_f decreases from left to right. Recall appears to decrease when N_f decreases because less common features lead to ineffective identification of related documents (smaller *true positive*) and a larger number of false negative (larger *false negative*).

In terms of purity, rand index, and precision, there exists an inflection point around $N_f = 1000$, where optimal clustering results were achieved. With a large feature space (e.g., with 30,000 features for WebKB), unrelated documents were likely to be grouped together because of irrelevant common terms, leading to a large number of false positive (hence lower precision). Some degree of feature removal reduced the amount of *noise* in the feature space and improved clustering effectiveness. It was shown in research that using various feature selection methods to eliminate up to 90% of term features resulted in either no loss or improvement of clustering and categorization accuracy [21, 33]. Further feature reduction from the inflection point degraded clustering performance when there were insufficient features for accurate document representation. As shown in Figures 3 and

4, similar patterns about the influence of feature selection were found with 20Newsgroup and RCV1 data.

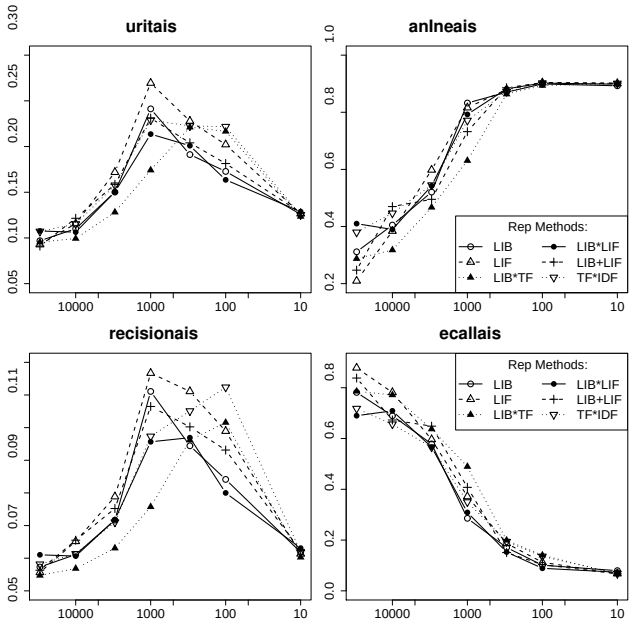

Figure 3: 20Newsgroup: Impact of feature selection

Interestingly, dramatically reducing the number of features from the inflection point only degraded *rand index* to a certain degree. We suspect that with an extremely small number of features to represent a large, diverse collection, clustering did not perform effectively. Having a good *rand index* score with only 10 features, for example in Figure 3 on the 20 Newsgroups data, is against our intuition about the situation. This raises questions about the effectiveness of *rand index* in evaluating clustering results and the circumstances under which *true negative* should or should not be considered.

4.6 HAC Clustering Results

The main experiments above were conducted using k-means clustering. In this section we present results from hierarchical agglomerative clustering (HAC) using complete link and an arc cosine distance function. From Tables 6, 7, 8, and 9, we have found results consistent with those from k-means clustering.

Table 6: WebKB with hierarchical clustering

Method	Purity	RIndex	Prec.	Recall	F_1
TF*IDF	0.459	0.623	0.283	0.235	0.253
LIB	*0.464*	*0.628*	*0.295*	0.233	*0.259*
LIF	0.453	0.484	0.239	**0.391**∗	**0.296**∗
LIB*TF	0.452	0.623	0.275	0.223	0.244
LIB+LIF	*0.473*∗	**0.641**	**0.308**∗	0.228	*0.261*
LIB*LIF	**0.485**∗	*0.633*	*0.301*∗	0.226	*0.257*

With WebKB and 20 Newsgroups data (shown in Tables 6 and 7), for example, LIB+LIF and LIB*LIF continued to perform competitively in terms of within-cluster accuracy

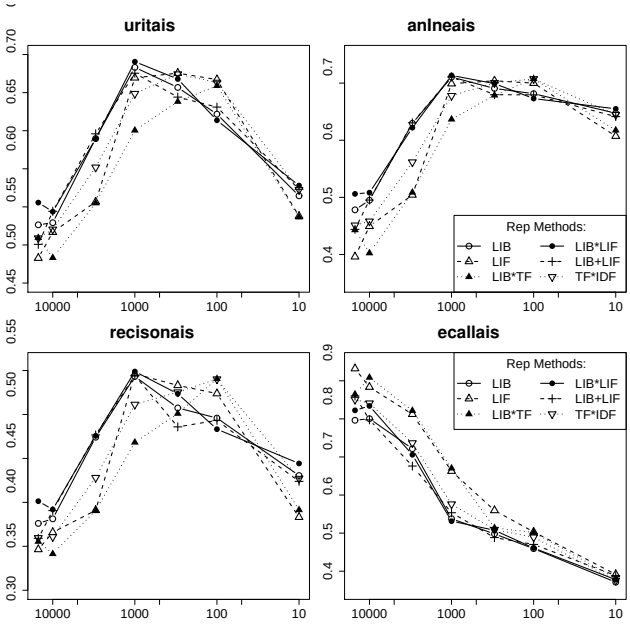

Figure 4: RCV1: Impact of feature selection

Table 7: 20 Newsgroup with hierarchical clustering

Method	Purity	RIndex	Prec.	Recall	F_1
TF*IDF	0.192	0.584	0.0582	0.475	0.103
LIB	0.183	*0.614*	0.0575	0.429	0.101
LIF	**0.240**∗	*0.891*∗	*0.0987*∗	0.142	*0.116*∗
LIB*TF	0.172	0.568	0.0567	**0.483**	0.101
LIB+LIF	*0.232*∗	*0.886*∗	**0.0988**∗	0.154	**0.120**∗
LIB*LIF	*0.224*∗	**0.892**∗	*0.096*∗	0.134	*0.111*∗

metrics such as purity and precision. LIF achieved better recall and F_1 than TF*IDF did on WebKB (Table 6).

Table 8: RCV1 with hierarchical clustering

Method	Purity	RIndex	Prec.	Recall	F_1
TF*IDF	0.448	0.357	0.312	0.865	0.458
LIB	*0.477*∗	*0.384*∗	*0.324*∗	0.864	**0.470**∗
LIF	**0.542**∗	**0.622**∗	**0.398**∗	0.391	0.392
LIB*TF	*0.464*∗	*0.370*	*0.319*∗	**0.875**	*0.467*∗
LIB+LIF	*0.489*∗	*0.399*∗	*0.325*∗	0.829	*0.466*
LIB*LIF	*0.492*∗	*0.399*∗	*0.322*∗	0.804	*0.458*

With 20 Newsgroups, RCV1, and NYTimes data, LIF consistently outperformed TF*IDF in multiple metrics (see Tables 7, 8, and 9). With the RCV1 corpus, all proposed methods produced significantly higher scores in purity, rand index, and precision than TF*IDF did.

5. DISCUSSION AND RELATED WORK

In the various experiments presented here, the proposed term weighting methods based on *least information* modeling performed very strongly compared to TF*IDF. In all experiments on the four benchmark collections, top perfor-

Table 9: NYTimes with hierarchical Clustering

Method	Purity	RIndex	Prec.	Recall	F_1
TF*IDF	0.840	0.709	0.785	0.511	0.586
LIB	*0.840*	0.693	0.777	0.489	0.564
LIF	**0.867**∗	**0.741**∗	**0.845**∗	0.501	*0.616*∗
LIB*TF	0.839	0.699	*0.788*	0.483	0.567
LIB+LIF	*0.844*	0.700	*0.787*	0.490	0.570
LIB*LIF	0.839	*0.728*∗	0.780	**0.544**∗	**0.617**∗

mance scores were achieved among the proposed methods. Whereas LIF well supported *recall*, LIB*LIF was overall the best method in the experiments and consistently outperformed TF*IDF by a significant margin, particularly in terms of purity, precision, and rand index. Although detailed results varied, the general observations from k-means clustering and hierarchical clustering were consistent.

Experiments showed that methods with the LIB quantity were more effective in terms of within-cluster accuracy (e.g., precision and purity). By emphasizing the discriminative power (specificity) of a term, LIB reduces weights of terms commonly shared by unrelated documents, leading to fewer of these documents being grouped together (smaller *false positive* and higher precision). LIF, on the other hand, helped to boost recall with the integration of term frequency. The different strengths of LIB and LIF indicate that they can be combined or used separately to serve various clustering purposes.

An additional interesting finding in this study is the inflection point in the clustering performance vs. # features plots. In various data and experimental settings, optimal clustering performance was achieved with 1000 features (selected by DF thresholding). Increasing or decreasing the number of features from the inflection point degraded clustering effectiveness in terms of purity, rand index, and precision. While similar patterns were observed in text clustering and categorization research, further investigation is needed to understand factors related to this phenomenon.

5.1 Related Models

The LIB*LIF scheme is similar in spirit to TF*IDF. By modeling (binary) term occurrences in a document vs. in any random document from the collection, LIB integrates the document frequency (DF) component in the quantity. LIF, on the other hand, models term frequency/probability distributions and can be seen as a new approach to TF normalization. Despite the similarity, our experiments showed LIB*LIF, based on the new *least information* formulation, were more effective than TF*IDF for document representation in the text clustering context. Least information modeling can be applied to other important processes such as information retrieval ranking, for which TF*IDF and its BM25 variation have produced strong empirical results.

Several works have attempted to justify clasic IDF from an information-theoretic view but have not established a direct, concrete connection [26]. Further development of notions around information-theoretic entropy led to findings such as *maximum entropy* and *minimum (mutual) information* principles, providing important guidance to inferential statistics for retrieval and evaluation [13, 31, 16, 4].

It has been shown that a term's IDF is the mutual information between the term and the document collection [30]. Mutual information is an application of *relative entropy* (KL divergence) that quantifies the difference between the joint probabilities and product probabilities of two random variables [8]. In the light of language modeling, IDF is equivalent to Kullback-Leibler (KL) divergence (relative entropy) between term probability distribution in a document and in the entire collection [1, 18].

KL divergence (relative entropy) measures discrimination information between two probability distributions by quantifying the entropy change in an asymmetric manner [18]. The asymmetry of KL divergence is due to the fact that it works in a directed manner to quantify bits needed to code samples of one distribution based on another. The proposed least information theory (LIT) offers a symmetric function and can be used as a distance measure. In addition, whereas KL is infinite given extreme probabilities (e.g., for rare terms), the amount of least information is bounded by the number of inferences.

5.2 Novelty and Significance

The least information theory is a new extension of Shannon entropy based on integration of micro uncertainty changes over a shifted probability distribution. LIB and LIF represent our initial attempt to apply LIT in modeling document representation for text clustering, which offers promising baseline results. The presented models may be further improved, for example, by integrating various randomness models and probability estimators such as those in [2]. LIT provides a new information quantity with which more sophisticated models can be proposed. Its implications will be beyond text clustering.

While the LIB and LIF models bear some resemblance to existing approaches such as divergence from randomness in [2] and information-based models in [7], the LIT formulation is fundamentally different. It is true that the common idea is to weight terms by measuring information they contribute, relative to signals in the collection/randomness. The key lies in how information is measured. LIT quantifies variances in $p(1 - \ln p)$, where p denotes the probability of an inference (e.g., that a term appears). Although this can be related to classic quantities such as $-\ln p$ (IDF-related), $-p \ln p$ (entropy-related), and $p \ln p/q$ (KL-related) commonly used in the literature, least information is not an arbitrary combination of existing quantities but a derivation based on close examination of expected characteristics. To our knowledge, no existing information theory or models share the major characteristics (see section 2.3) of the presented least information quantity (in equation 9).

6. CONCLUSION

We presented the *least information* theory (LIT), which quantifies information in varied probability distributions. We observed several important characteristics of the proposed information quantity. Two basic quantities were derived from the theory for term weighting and document representation, which we used separately and combined in fusion methods for document clustering.

Research was conducted to evaluate the effectiveness of proposed methods compared to TF*IDF, which had been extensively used in text clustering research. Experiments on several benchmark collections showed very strong per-

formances of LIT-based term weighting schemes. In most experiments, the proposed methods, especially LIB*LIF fusion, significantly outperformed TF*IDF in terms of several evaluation metrics.

While we have demonstrated superior effectiveness of the proposed methods, the main contribution is not about improvement over TF*IDF. Of greater significance is the new approach to information measurement and term weighting based on the *least information* theory (LIT), which enables a different way of thinking and provides a new information measure for modeling various information processes.

Acknowledgment

This research is partially supported by IMLS grant #LG-06-11-0332-11. We thank Xiaoli Song and anonymous JCDL'13 reviewers for constructive comments.

7. REFERENCES

[1] A. Aizawa. The feature quantity: an information theoretic perspective of TFIDF-like measures. In *SIGIR'00*, pages 104–111, 2000.

[2] G. Amati and C. J. van Rijsbergen. Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Transactions on Information Systems*, 20(4):357–389, October 2002.

[3] D. Arthur and S. Vassilvitskii. k-means++: the advantages of carefull seeding. In *SIAM'07*, pages 1027–1035, 2007.

[4] J. A. Aslam, E. Yilmaz, and V. Pavlu. The maximum entropy method for analyzing retrieval measures. In *SIGIR'05*, pages 27–34, 2005.

[5] R. Baierlein. *Atoms and Information Theory: An Introduction to Statistical Mechanics*. W.H. Freeman and Company, 1971.

[6] M. W. Berry. *Survey of text mining: clustering, classification, and retrieval*. Springer, 2004.

[7] S. Clinchant and E. Gaussier. Information-based models for Ad Hoc IR. In *SIGIR'11*, pages 234–241, 2011.

[8] T. M. Cover and J. A. Thomas. *Entropy, Relative Entropy and Mutual Information*, pages 12–49. John Wiley & Sons, 1991.

[9] M. Craven, D. DiPasquo, D. Freitag, A. McCallum, T. Mitchell, K. Nigam, and S. Slattery. Learning to extract symbolic knowledge from the World Wide Web. In *AAAI'98*, pages 509–516, 1998.

[10] J. D. Fast. *Entropy: the Significance of the Concept of Entropy and Its Applications in Science and Technology*. McGraw-Hill, 1962.

[11] C. Fox. *Information and misinformation: an investigation of the notions of information, misinformation, informing, and misinforming*. Contributions in librarianship and information science. Greenwood Press, 1983.

[12] A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: a review. *ACM Comput. Surv.*, 31(3):264–323, Sept. 1999.

[13] E. T. Jaynes. Information theory and statistical mechanics. ii. *Phys. Rev.*, 108:171–190, Oct 1957.

[14] X. Ji and W. Xu. Document clustering with prior knowledge. In *SIGIR'06*, pages 405–412, 1996.

[15] K. S. Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 60:493–502, 2004.

[16] P. B. Kantor and J. J. Lee. The maximum entropy principle in information retrieval. In *SIGIR'86*, pages 269–274, 1986.

[17] S. Kullback. Letters to the editor: The Kullback-Leibler distance. *The American Statistician*, 41(4):338–341, Nov 1987.

[18] S. Kullback and R. A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22:79–86, 1951.

[19] K. Lang. Newsweeder: Learning to filter netnews. In *Proceedings of the Twelfth International Conference on Machine Learning*, pages 331–339, 1995.

[20] D. D. Lewis, Y. Yang, T. G. Rose, and F. Li. RCV1: A new benchmark collection for text categorization research. *J. Mach. Learn. Res.*, 5:361–397, 2004.

[21] T. Liu, S. Liu, Z. Cheng, and W.-Y. Ma. An evaluation on feature selection for text clustering. In *ICML'03*, 2003.

[22] J. B. Lovins. Development of a stemming algorithm. *Mechanical Translation and Computational Linguistics*, 11:22–31, 1968.

[23] D. M. MacKay. *Information, Mechanism and Meaning*. The M.I.T. Press, 1969.

[24] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[25] A. Rapoport. What is information? *ETC: a review of general semantics*, 10(4):5–12, 1953.

[26] S. Robertson. Understanding inverse document frequency: on theoretical arguments for IDF. *Journal of Documentation*, 60:503–520, 2004.

[27] S. Robertson and H. Zaragoza. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends® in Information Retrieva*, 3(4):333–389, 2009.

[28] E. Sandhaus. The New York Times annotated corpus. Linguistic Data Consortium, 2008.

[29] C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423 and 623–656, July and October 1948.

[30] M. Siegler and M. Witbrock. Improving the suitability of imperfect transcriptions for information retrieval from spoken documents. In *ICASSP'99*, pages 505–508. IEEE Press, 1999.

[31] F. Snickars and J. W. Weibull. A minimum information principle: Theory and practice. *Regional Science and Urban Economics*, 7(1):137–168, 1977.

[32] I. H. Witten and E. Frank. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, San Francisco, 2nd edition, 2005.

[33] Y. Yang and J. O. Pedersen. A comparative study on feature selection in text categorization. In *ICML'97*, pages 412–420, 1997.

[34] D. Zhang, J. Wang, and L. Si. Document clustering with universum. In *SIGIR'11*, pages 873–882, 2011.

[35] Y. Zhao and G. Karypis. Evaluation of hierarchical clustering algorithms for document datasets. In *CIKM'02*, pages 515–524, 2002.