How Informative are the Concept Associations Discovered by a Document Cluster?

Abstract. Over the past two decades, document clustering has matured into a fundamental data mining technique, with potential for use in diverse areas in industry. In this paper, we address an aspect of the structure of document clusters: namely, the information captured in concept associations that are discovered by document clustering. Can we *measure* this information in a principled and statistically sound manner? This problem is germane to several applications of data mining to industry, including eDiscovery and homeland security.

We approach this problem as follows. First, we construct representations for concepts in document clusters using frequent itemsets. We then use these representations to frame the notion of associations between concepts. These are the associations that the document cluster has "discovered." Next, we use information theory to measure the information contained in these concept associations. This requires us to measure what is the prior probability of such an association occurring, based on the statistics of the corpus. Information theory tells us that the more unlikely this association, the more informative it is.

Our approach results in two information theoretic scores for each document cluster. These scores are interpretable: the user can be informed why a cluster has a certain score. We also show how to implement our techniques with very little overhead by using structures that are already computed during document clustering. Finally, we demonstrate the flexibility of our framework by using it to perform a variety of tasks centered around concept associations and their information content on benchmark datasets for document clustering.

1 Introduction

When a clustering algorithm groups documents into clusters, it is associating concepts that occur in these documents. The clustering algorithm can be said to have discovered these concept associations. There has been a great deal of research on better algorithms: faster, producing tighter clusters (as measured by, say, the value of the objective function being minimized). There has also been research on convergence properties of clustering algorithms. However, there is relatively lesser research on measuring, in a statistically sound manner, the *amount of structural information* contained in a document cluster. In this paper, we address the following specific question on structural information:

How do we measure the information contained in the conceptual associations discovered by a document cluster?

For the rest of the paper, we use information in this specific sense: namely, informativeness will mean structural informativeness of concept associations in a document cluster.

The impetus for our work came from applications of clustering to a market leading commercial information management product. Vendors of such products develop solutions incorporating document clustering that are tailored towards commercially significant applications. We outline pertinent challenges in two such example applications below.

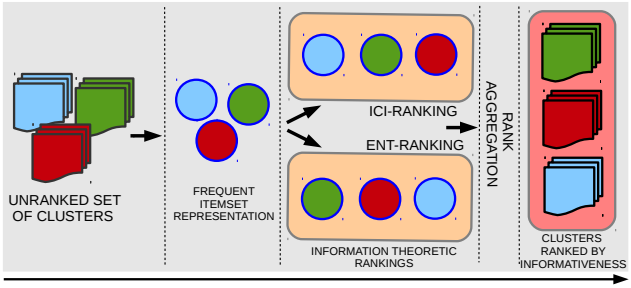


Fig. 1. An overview of the FICS algorithmic framework to rank document clusters by the informativeness of their concept associations.

eDiscovery is a significant user of data mining and information retrieval technologies in large enterprises.[1] Document review is estimated to be between 60% and 80% of eDiscovery cost (21). Therefore, vendors increasingly provide clustering to aid document review. A user interacting with a cluster-based eDiscovery system is often not specifically looking to answer a query in early stages of the review process, but rather wants to uncover *informative unexpected associations*. In addition, such informative associations must be discovered *quickly*, since the early case assessment (ECA) period[2] is often only a few weeks, and millions of documents may be

[1] Market in 2012 $1.4B, expected market in 2017 $2.9B (10).

[2] The stages of eDiscovery are explained at www.edrm.net. ECA is an early stage, where the organization must decide its liability and whether to settle out of court or go to trial.

under scrutiny. This makes a manual inspection for informative associations infeasible. It would be useful if informative document clusters were tagged up front, and examined early since they could potentially change the nature of the litigation.

Homeland security is a significant emerging application for data mining (22). Homeland security applications that use document clustering typically throw up many associations that are typical and not informative. This glut of non-informative associations is particularly bothersome given the premium and time-criticality of analyst resources in homeland security applications. Homeland security agencies need techniques that can identify hidden and unexpected concept associations for immediate attention.

Note that in both our motivating examples, the informativeness measures used must be *interpretable* by the end-user. Namely, the user should be able to readily understand why a particular cluster is more informative than another.

Accordingly, we formulated our research goal as follows:

> Derive a mathematical framework in which the information contained in conceptual associations discovered by document clustering can be measured. The framework should be interpretable.

Our approach is guided by the following statistical ideas.

1. Information theory quantifies precisely the amount of information contained in an event. If the event has probability p, then the information contained in its occurrence is $-\log p$.
2. Therefore, in order to measure the information content in a cluster, we need to (a) represent concepts and their associations discovered by the cluster and (b) measure their probability of occurrence.
3. We measure the probability of occurrence of a concept association, with respect to the "baseline" statistics of the corpus.
4. An association between two important concepts is more informative than one between an important concept and an unimportant one. This can be formalized using the notion of information-theoretic entropy.

An overview of our approach is illustrated in Fig. 1.

We use suitable notions of frequent itemsets to represent concepts. We call our algorithmic framework FICS, for "frequent itemset connection scoring."[3] FICS computes two information theoretic scores for each cluster. These scores measure, in an objective manner, the informativeness of the cluster.

The algorithmic framework of FICS was designed with the following system-level advantages in mind.

- FICS can be computed with minimal overhead on top of clustering since it uses data that is already computed in most document clustering.
- FICS does not require any external information such as wiki, etc.
- FICS does not require administrator input or judgements.
- FICS is readily interpretable.

Due to these advantages, FICS can co-exist with a very small footprint in any clustering system.

1.1 Our Contributions

Our contributions are outlined at a high-level below.

1. We motivate the need for structural informativeness measures for document clusters, using real-world applications in enterprise information management having high commercial impact (§1).
2. We build a mathematical framework (§4 through §6) that inspects the concept structure in a cluster, associates to such a structure the objective measures of informativeness that we seek.
3. We provide experimental usage templates of the above framework that model real-world applications (§8).

2 Related Work

Our work has conceptual linkages to three streams of research.

2.1 Interestingness Measures for Association Rules Mining

Bearing in mind the volume of the work on interestingness for association rules, and the indirect link to our work, we will restrict ourselves to pointing out the excellent surveys that have already been written on this mature field.

Broadly speaking, there are two approaches to developing interestingness notions for association rules: *objective*, based on frequencies of occurrences (see, for example, (3; 20)) and *subjective*, based on user beliefs (for example, (18; 23)). (19) also provides a subjective notion of interestingness by quantifying unexpectedness in terms of user beliefs. This notion of interestingness is similar to our notion of informativeness, but is subjective as opposed to ours, which is objective.

Following the initial papers on both of these approaches, there was a flurry of activity in the field. The first survey to appear—(13)—reviews 17 measures, both subjective and objective. A conceptual introduction to interestingness notions is provided by (4). Further research activity led to two more surveys and comparative analyses. (25) (see also (26)) focus on the appropriate choice of interestingness measures, and compare 21 such measures. (17) use the context of data mining workflows, and discuss user intervention in it. Another notable survey is (11), who offer a broader review, stepping out of the data mining context strictly, and focus on probabilistic measures. They review 38 objective measures, along with 2 subjective ones, culling from 60 references. A recent review—(15)—categorizes interestingness measures into syntactical and probabilistic approaches.

[3] Also "find informative clusters."

At this stage, we examine the linkages to our work at the conceptual level. Association rules are composed of implication expressions of the form $X \rightarrow Y$, where X and Y are itemsets in a database. A famous (but perhaps apocryphal) example is $\{diapers\} \rightarrow \{beer\}$. Two measures of rule evaluation appear very early in the literature: (a) Support (of the rule) (2), defined as the fraction of transactions that contain both X and Y; (b) Confidence (3), defined as the fraction of the transactions containing X that also contain Y. (1; 24) pointed out the shortcoming of the confidence measure: it does not take into account the baseline frequency of Y. This led to the introduction of Lift (14) (also known as Interest (5)), defined as the ratio of the support to that which would be expected had X and Y been independently occurring in the database.

There is a fundamental difference between the aforementioned notions of interestingness in association rules and our own notion of informativeness. The aforementioned notions, very roughly, quantify whether or not the association under consideration actually exists, or is a spurious association unearthed by the algorithm. Support indicates the significance of a rule in the database. Rules with very low support values represent very small numbers of transactions: namely, these are outliers that are not likely to be profitable from a commercial data mining point of view. Similarly, the confidence measure, which gives rise to several other measures, is a confidence that the rule actually exists in the database. In our application, we are assuming that the clusters that have been discovered by the clustering algorithm are valid. *Assuming their validity*, which are the most informative in the specific sense of information contained in their concept associations? This is the question we address.

2.2 Frequent Itemsets in Document Clustering

We also mention some works that use frequent itemsets in the context of text clusters, but have very different goals from our work. The first, (9), uses frequent itemsets to reduce the dimensionality of the text clustering problem, and to perform clustering. The second, (16), also seeks to produce good clusters using itemsets, but consider "closed interesting" itemsets (rather than frequent itemsets). The goals of both the aforementioned works is to produce clusters. Our goal is not to produce clusters, but to measure the information content contained in concept associations in clusters that are already formed by other means. Finally, (7) associate to a document cluster a graphical "shape" using frequent itemsets extracted from the cluster. They do not consider the problem of measuring information content.

2.3 Validity of Clusters

We use this section partly to clarify the scope of our work. We assume that our clusters are valid. There is, of course, a vast amount of literature on the question of ascertaining when clusters are valid. See, for instance, (12) for a review, and (8) for information-theoretic external measures. However, as noted, these are outside the scope of our work. We assume that the user has, separately, filtered out those clusters that are invalid, using some means at their disposal. We should also point out that a standard way to determine whether a document cluster is valid is by means of its internal similarity (ISim, see Sec. 4), and by this token the clusters in our experimentation section are all valid.

3 Key Ideas

For the purposes of the following discussion, a *concept* is a set of terms that frequently co-occur within documents of the document corpus. We will adopt a concept-centric view of clustering.

Idea 1: Differentiate between Local and Non-local Concept Associations

Consider a set of documents placed into a cluster by a clustering algorithm. We may also think of this as a mapping of the concepts contained in the documents to the cluster. Take for example a cluster on US politics. Consider the concept represented by the words "GOP, republican, election," which frequently co-occur in documents on politics. The clustering algorithm will likely map this concept to the cluster on US politics. In this sense, the clustering algorithm has mapped a concept to a specific cluster. By doing so, it has grouped together documents that contain a co-occurrence of these three terms (this latter statement is the document-centric version). We call such an association of concepts to clusters **local** associations.

That is not all that a clustering algorithm does. For example, if another set has co-occurrences of "republican national convention, running mate, Tampa FL, election," then the clustering algorithm will likely place such documents into the same cluster on US politics as the previous set of documents. At this stage, the algorithm has *associated two different sets of co-occurrences of words*: it has discovered that the two sets "GOP, republican, election" and "republican national convention, running mate, Tampa FL, election" should be associated with each other. We informally call such associations of concepts to each other **non-local** associations.

Let us differentiate the two types of associations. Grouping together documents containing "GOP, republican, election" is easy: a simple "grep" command would uncover such an association. On the other hand, the association *across* two sets of concepts would not have been easily known without the clustering algorithm. Namely, non-local associations are harder to discover because they represent interactions "at a distance" within the data. Our formal name for non-local associations is *connections* since they connect different concepts in the data through the process of clustering. The key now is to measure the information contained in these such connections in a statistically sound manner.

Idea 2: Unexpected Connections are Informative

In the previous section, we identified a structural source of information in a cluster: namely, the connections between concepts in the data that the cluster has captured. In this section, we describe our second idea. We return to our example to illustrate it. The term "election" was common between both the concepts that the clustering algorithm associated. In other words, it is the connection made through the term "election" that allowed the clustering algorithm to associate the two concepts. How hard was it to find this connection? Intuitively, if the term "election" is frequent, then it is easy to find this connection, and therefore associate the two concepts. If, on the other hand, the term "election" is infrequent, then finding the connection is harder.

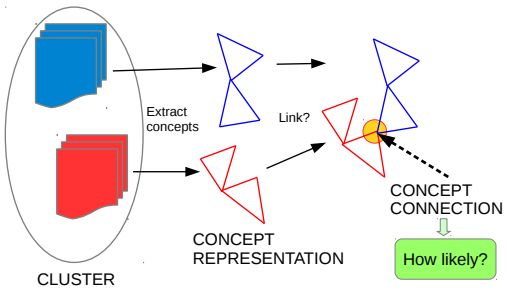


Fig. 2. Frequent itemsets extracted from sets of documents represent their concepts. The linkages between such concepts form connections (non-local associations). Connections make the cluster informative. The less likely a connection, the more information it contains. Discussion in Sec. 3.

We make this intuition precise using information theoretic arguments. Information theory allows us to say precisely that "the more likely a connection between two concepts is, the less informative it is. Conversely, an unlikely connection between concepts is more informative." The discussion thus far is illustrated in Fig. 2.

Idea 3: Connections Between Equally Important Concepts are More Informative

When a cluster finds an association between two concepts, the informativeness of this association also depends upon the relative importance of the two concepts. Consider an importance value for each concept. Then, for a given sum of importance values, an association between two concepts of equal importance is more informative than one between an important concept and an unimportant one. Once more, information theory provides us with a framework which makes this statement precise and quantitative. The information theoretic notion that captures this is the entropy, which is highest when a probability distribution has equal probability masses on each of its states.

At this stage, we are ready to start building the quantitative framework within which these ideas can be made precise, and computations carried out. We begin with preliminaries in Sec. 4, build the components of our FICS framework in Sec. 5, and finally present the FICS algorithm in Sec. 6.

4 Preliminaries

Our data $\mathcal{D}$ will be comprised of a document collection (or corpus) $\mathcal{D}=\{D_1,...,D_n\}$. A *$K$-clustering* $\mathcal{C}=\{C_1,...,C_K\}$ of $\mathcal{D}$ is a partition of $\mathcal{D}$ into K disjoint non-empty subsets whose union is $\mathcal{D}$. We assume a standard preprocessing of documents for clustering: tokenization of words, stemming, removal of stoplists, and removal of words occurring very infrequently (less than thrice) in the corpus. The resulting set of tokens will be called *terms*. The *vocabulary* of $\mathcal{D}$ is the set of all terms in the documents of $\mathcal{D}$. Sets of terms will be referred to as itemsets. Itemsets will be denoted by F, with subscripts when necessary. The *support* of an itemset F in $\mathcal{D}$, denoted by $\text{supp}(F,\mathcal{D})$, is the set of all documents in $\mathcal{D}$ in which all the terms of F co-occur. Likewise, $\text{supp}(F,C)$ is the set of all documents in the cluster C that contain all the terms of F.

For the definitions below, C is a generic cluster in $\mathcal{C}$. First we define notions of frequent, and frequent relative to C.

Definition 1. *1. Let α be a positive integer that is called* minimum support. *F is said to be* frequent *(w.r.t. α) if the size of the support of F exceeds α.*
2. A frequent itemset F is said to be frequent relative to C *if $\text{supp}(F,C)>\beta\text{supp}(F,\mathcal{D})$, where β is a threshold and $0<\beta<1$.*
3. Then F is said to be maximal frequent relative to C *if it is the maximal (w.r.t. set inclusion) itemset with the preceding property.*

By "frequent," we henceforth mean "maximal frequent relative to a cluster C."

For a term t that occurs in C, the *L_1 norm of t with respect to C*, denoted $\text{norm}(t)$, is defined as the coefficient of t in the centroid of C when the centroid has been normalized so that its coefficients sum up to 100.

We denote by $C^{[\ell]}$ the ℓ most frequent terms in C. Equivalently, these are the ℓ terms having the highest L_1 norm in C.

We denote the maximal frequent itemsets of our generic cluster C by $\mathscr{F}(C)=\{F_1,...,F_q\}$. These represent the set of concepts in the cluster. However, in order to reduce computational complexity, we work with a more manageable set of frequent itemsets—those whose terms are all taken from $C^{[\ell]}$.

Definition 2. *We define $\mathscr{F}(C^{[\ell]})$ as the set of frequent itemsets for C such that for all $F\in\mathscr{F}(C^{[\ell]})$, each term in F is from $C^{[\ell]}$.*

Finally, a measure of cluster cohesiveness that we will refer to is ISim (internal similarity), defined as the average cosine distance of the documents in a cluster from the cluster centroid.

5 Components of FICS

In this section, we develop the components required for our algorithm. The algorithm itself is presented in Sec. 6. Our setting throughout this section will be a generic cluster C in a K-way clustering of $\mathcal{D}$.

5.1 Cluster Representation

First, we explain the information needed by our algorithm about each cluster. This information is encapsulated in the cluster representation, defined below.

Definition 3. *The* cluster representation *of a cluster C, denoted $\mathcal{R}(C)$, is comprised of the following data:*

1. The terms in $C^{[\ell]}$, along with their L_1 norms.

2. *The maximal frequent itemsets $\mathscr{F}(C^{[\ell]})$. These represent the concepts in the cluster.*
3. *An inverted index for terms in (1) and their occurrences in C, or the term-document matrix for terms in (1) restricted to documents in C.*

Significance of ℓ The choice of ℓ determines the user's threshold for significant associations between concepts. In other words, by choosing ℓ, the user specifies the number of terms that they think should be considered when computing the information-theoretic measures described in the following sections. By specifying a value for ℓ, the user is saying that associations centered around terms that occur at coordinates ranked lower than ℓ in the centroid vector (where the ranking is by magnitude), are to be ignored. Therefore, a higher value for ℓ means that the user wants a more conservative definition for concepts and their associations. The examples in Sec. 8 will make this clear.

We have experimented with various values of ℓ and recommend $\ell=10$ or 15. This is consistent with literature in both clustering and information retrieval where the top 10 to 20 terms (usually ranked by information gain) are deemed to be a good description of the cluster.

5.2 Facet Connection Information

With our cluster's concise representation fixed, we are ready to associate to each such representation a quantitative "level of informativeness." We will need some more definitions.

Itemset Connections and Facets

Definition 4. *Consider a pair of distinct maximal frequent itemsets (F_i,F_j) from $\mathscr{F}(C^{[\ell]})$.*

(i) The connection *between F_i and F_j is defined as*

$$\chi_{i,j}=F_i\cap F_j. \tag{1}$$

(ii) When $\chi_{i,j}\neq\emptyset$, (F_i,F_j) is called a facet *of C. The set of facets of C will be denoted by* $\mathsf{Facets}(C)$.

Facet Connection Likelihood Let $(F_i,F_j)\in\mathsf{Facets}(C)$ and $\chi_{i,j}$ be the corresponding connection.

Definition 5. *The* likelihood *of the facet connection between F_i and F_j is defined as*

$$P(\chi_{i,j})=\frac{\sum\{\mathrm{norm}(t)\colon t\in\chi_{i,j}\}}{\sum\{\mathrm{norm}(t)\colon t\in F_i\cup F_j\}} \tag{2}$$

Proposition 1. *Hypothesis as in Def. 5. Then $P(\chi_{i,j})$ is the likelihood of the connection between F_1 and F_2 being $\chi_{i,j}$, given the statistics of the corpus $\mathcal{D}$, assuming an independence model for document generation.*

Proof. Place a uniform measure on the documents in the cluster. The independence model for documents implies that features occur independently of each other. Then the numerator represents the measure of documents in C that contain the terms common to F_1 and F_2. The total measure of documents in C that contain any of the terms in $F_1\cup F_2$ is given by the denominator.

Notice that if each term was weighted equally, then $P(\chi_{i,j})$ is just the Jaccard coefficient between F_i and F_j, considered purely as sets.

Facet Connection Information Next, we measure the information content in the statement "F_1 and F_2 have $F_1\cap F_2$ in common." Information theory tells us that the information content of a statement is the negative logarithm of the probability of that statement (6). In this case, it is the negative logarithm of the probability of the proposition $(\chi_{i,j}\subset F_1)\wedge(\chi_{i,j}\subset F_2)$. This gives us the following definition.

Definition 6. *The* facet connection information, *denoted $\mathsf{FCI}(F_i,F_j)$, of the facet (F_i,F_j) (or of their connection $\chi_{i,j}$) is defined as*

$$\mathsf{FCI}(F_i,F_j)=-\log[P(\chi_{i,j})]. \tag{3}$$

5.3 Facet Entropy

We now turn our attention to the distribution of weights between the two itemsets that form the facet under consideration. The more equal this distribution, the more entropy the facet possesses.

Facet Proportions

Definition 7. *Let $(F_i,F_j)\in\mathsf{Facets}(C)$. Define the proportion*

$$p_i=\frac{\sum\{\mathrm{norm}(t)\colon t\in F_i\}}{\sum\{\mathrm{norm}(t)\colon t\in F_i\}+\sum\{\mathrm{norm}(t)\colon t\in F_j\}}, \tag{4}$$

and the proportion p_j similarly.

Facet Entropy Namely, we wish to capture the information in the statement "the total weight of a facet is divided between the two constituent itemsets in the proportion $p_i\colon p_j$."

Definition 8. *The* facet entropy *of (F_i,F_j), denoted by $\mathsf{ENT}(F_i,F_j)$, is defined as*

$$\mathsf{ENT}(F_i,F_j)=-[p_i\log(p_i)+p_j\log(p_j)]. \tag{5}$$

5.4 Subroutine Compute_FCI_ENT

We put the preceding components into a subroutine that can be called for a cluster C.

In summary, the framework built thus far gives a quantification to the idea that "an unexpected connection between two concepts is informative. It is all the more informative if the two concepts are equally important."

Algorithm 1: Subroutine to compute FCIs and ENTs for all the facets in C.

Data: Representation $\mathcal{R}(C)$ of Cluster C
Result: List of FCIs and ENTs of all the facets in C
Compute $\mathsf{Facets}(C)$;
for $(F_i,F_j) \in \mathsf{Facets}(C)$ **do**
  Compute $\chi_{i,j}$ and (p_i,p_j);
  Compute $\mathsf{FCI}(F_i,F_j)$;
  Compute $\mathsf{ENT}(F_i,F_j)$;
end
return $\{(\mathsf{FCI}(F_i,F_j),\mathsf{ENT}(F_i,F_j)) : (F_i,F_j) \in \mathsf{Facets}(C)\}$

6 The FICS Algorithm

We now gather all the components defined in the previous section into a single algorithmic framework that we call FICS. FICS is a flexible framwork, and can be used in several ways. Accordingly, the final two steps (described below) can be implemented in multiple ways; we only show the simplest for clarity.

Thus far, we have computed information theoretic scores for facets. However, there are, in general, multiple facets in a single cluster. Therefore, a simple way to give a single score to a cluster is to take the maximum of its facet scores.

Definition 9. *The* cluster FCI *for* C*, denoted by* $\mathsf{FCI}(C)$*, is defined as*

$$\mathsf{FCI}(C) = \max\{\mathsf{FCI}(F_i,F_j) : (F_i,F_j) \in \mathsf{Facets}(C)\} \tag{6}$$

Likewise, the cluster ENT *for* C*, denoted by* $\mathsf{ENT}(C)$*, is defined as*

$$\mathsf{ENT}(C) = \max\{\mathsf{ENT}(F_i,F_j) : (F_i,F_j) \in \mathsf{Facets}(C)\} \tag{7}$$

Finally, we need to aggregate ranks produced by $\mathsf{FCI}(\cdot)$ and $\mathsf{ENT}(\cdot)$. Let $L_{\mathsf{FCI}}(\mathcal{C})$ and $L_{\mathsf{ENT}}(\mathcal{C})$ be the two ranked lists of clusters. The choice of rank aggregation algorithm will be application specific. Sometimes, our goal is not to rank all the clusters, but only to find the most informative ones. In such cases, we may use the following simple scheme, which we call Top-r-Intersection. Starting with $r=1$, we inspect the clusters in the Top-r sub-lists of both lists $L_{\mathsf{FCI}}(\mathcal{C})$ and $L_{\mathsf{ENT}}(\mathcal{C})$. We increase r until there is a non-empty intersection of these two Top-r sub-lists. We return this cluster.

At this stage, we are ready to put everything together and describe the algorithm FICS.

Algorithm 2: The FICS algorithmic framework.

Data: Unordered Set of Document Clusters $\mathcal{C} = \{C_1,...,C_K\}$
Result: List of Document Clusters $L(\mathcal{C})$, Ranked by Informativeness
for i *in* $1,...,K$ **do**
  Compute $\mathcal{R}(C_i)$;
  Call $\mathsf{Compute\_FCI\_ENT}(\mathcal{R}(C_i))$;
  Compute $\mathsf{FCI}(C_i)$; Compute $\mathsf{ENT}(C_i)$;
end
Construct Ranked List $L_{\mathsf{FCI}}(\mathcal{C})$;
Construct Ranked List $L_{\mathsf{ENT}}(\mathcal{C})$;
Aggregate Ranked Lists to Construct $L(\mathcal{C})$;
return $L(\mathcal{C})$

We walk the reader through the pseudo-code of FICS (Algorithm 2). FICS obtains the set of document clusters computed by the clustering algorithm (such as K-means). It then computes, for each cluster, the representation of Def. 3, followed by a call to Compute_FCI_ENT which gets a list of (FCI, ENT) pairs for each facet in C. Next, it computes a single FCI and ENT for the cluster from this list. The body of this loop consists of lines 2-4.

At this point, FICS can construct two ranked lists: one ranked by descending order of FCI, and the other similarly by ENT. It finally aggregates these two ranked lists to produce a single ranked list that captures the informativeness of the clusters. Clusters at the top of this list are more informative than those at the bottom.

7 Implementation and Complexity

FICS has an implementation with very low overhead. This may seem surprising given that frequent itemsets can be worst-case exponential to compute. The key to having a constant time overhead for FICS is using only the terms in $C^{[\ell]}$ in the cluster representation $\mathcal{R}(C)$ (Def. 3).

Proposition 2. *FICS can be implemented to run in* $O(1)$ *for a fixed value of* K*, and is* $O(K)$ *otherwise.*

Proof. We observe that the representation $\mathcal{R}(C)$ of a cluster C can be obtained readily from the cluster centroid. It is here that our use of the L_1 norm in our computations, rather than a L_2 norm, allows for an economical implementation that simply "piggy-backs" on top of the centroid computation already performed by the clustering algorithm. Note that this choice was made possible because document vectors and centroids lie in the all-positive quadrant. Since many commonly used text clustering algorithms (such as K-means and its derivatives) already compute the centroid of each cluster at each iteration of the algorithm, this centroid is already available.

Once the cluster representation is available, there is only a fixed bounded amount of computation to be done, for both FCI and ENT computations. This is because we use only the terms in $C^{[\ell]}$ in the cluster representation $\mathcal{R}(C)$, instead of representing a cluster using all of its vocabulary.

In summary, FICS adds negligible overhead to the actual cost of clustering for the K-means family of clustering algorithms, since it uses structures that are already computed during the run of the K-means algorithm.

8 Experiments

Datasets. We experimented with three datasets that are widely used as benchmarks in document clustering. The first is the 20 Newsgroups dataset, denoted by `N20`, which contains roughly 20,000 articles posted to 20 usenet newsgroups. The articles are more or less evenly divided between the newsgroups; however some newsgroups are highly related, while others are not. The second and third datasets are subsets of `REU`, the Reuters-21758 dataset, collected and labelled by the Carnegie group and Reuters while developing the CONSTRUE system. There are 82 primary topics in the documents in this dataset. The R52 subset is obtained by considering only documents with a single topic, and retaining only those topics that have at least one test and one train document. This leaves only 52 of the original 90 classes, and has 9100 documents. We used the following two subsets of R52:

1. The train split of the 'R52' subset of `REU` which has 6532 documents.
2. The 'R8' subset obtained by taking the eight most frequent topics of R52. The resulting dataset has 7674 documents.

We take only the training documents of R52 in order to prevent a set-inclusion relationship between our two datasets.

There are multiple versions of these datasets. For replicability, we used the versions provided at http://web.ist.utl.pt/acardoso/datasets/. A summary of our datasets is provided in Table 8.

Dataset	Documents	Terms	Classes
`N20`	18821	91652	20
`REU` R52 Train	6532	16145	52
`REU` R8	7674	28140	8

Table 1. Datasets used

Protocol. We performed a K-way clustering of each dataset, which we denote by $\mathcal{D}$. For replicability, we used the repeat-bisect algorithm from the CLUTO toolkit,[4] rather than those which require random seeding. We varied K over a wide range:

$$K \in [5,10,15,20,25,30,35,40,50,60,70,80,90,100].$$

We chose $\alpha=15, \beta=0.3$ so as to gather a large number of maximal frequent itemsets for each cluster. We ran FICS on the resulting cluster representations. We retained all the intermediate information: for example, the computations of $\mathsf{FCI}(\cdot)$ and $\mathsf{ENT}(\cdot)$ for every facet of each cluster.

Since there are no available algorithms that measure information contained in concept associations of a cluster, our baseline is provided by the informal procedure that is employed today while inspecting clusters for their information. This is to our advantage, since it allows us to demonstrate how FICS would be used in real-world settings. We design our experiments accordingly: four of the six experiments are in the setting of a user session. The design at a high-level is explained in Table 2.

Experiment	Setting	Goal
1 and 2	Simulated User Session	Give examples of informative and non-informative clusters
3 and 4	Simulated User Session	Demonstrate flexibility of FICS using non-standard tasks
5 and 6	Analysis of Distributions	Study distributions of $\mathsf{FCI}(\cdot)$ and $\mathsf{ENT}(\cdot)$

Table 2. The design of our experiments, with high-level goals. Four of the experiments are intended to convey how FICS could be used for diverse tasks in a session with users (conducted by a data mining vendor, for example). The other two experiments study the distributions of information theoretic measures that FICS uses.

The first two experiments study clusters having few facets. In the first experiment, there are three facets, and in the second experiment, there are two.

[4] Available at http://glaros.dtc.umn.edu/gkhome/views/cluto.

8.1 Experiment 1: Use FICS to find a Cluster Capturing an Informative Facet

Objective To find a cluster with few facets that captures an informative facet in a given dataset, and explain to a user why the facet is informative.

Methodology We use the dataset N20 for this experiment. We cluster N20 into $K = 15$ clusters for a manageable list of clusters that can be examined by the user. For the set of clusters $\mathcal{C}$, we construct the two lists $L_{\mathsf{FCI}}(\mathcal{C})$ and $L_{\mathsf{ENT}}(\mathcal{C})$. Finally, we use the Top-r-Intersection subroutine in order to identify the most informative cluster with at most three facets.

Result A particular cluster that we will call C_{inf} appears first on both ranked lists, and is therefore deemed the most informative. One of its facets appears first on one list, second on the other; and the first position in the second list is also taken up by another of its facets. The cluster summary for C_{inf} is provided in Fig. 3 and Table 3(a).

Discussion Cluster C_{inf} brings together two different and long-running conflicts in the middle east. A user interested in the general theme of the cluster—conflicts in the middle east—and interested in the first conflict would probably want to know more about the second also. There are structural and geographical similarities between the two conflicts: these conflicts relate to each other although they are separated in time.

Now, let us discuss why the association between these two conflicts brought together by C_{inf} is informative. It is informative because it is statistically non-obvious. When we say non-obvious, we mean in the context of co-occurrence of these two conflicts in documents in our collection. To validate the scores that FICS has given to this cluster, we should verify that the facet captured in this cluster is unexpected, given the baseline statistics of co-occurrence of the dataset.

Indeed, we find that *the two conflicts rarely co-occur in documents in the collection.* For example, the terms "israel" and "armenian" rarely co-occur in a document. A user doing a sequence of "grep" operations on keywords obtained from the first conflict would not readily obtain documents on the second conflict. Therefore, it would be difficult to go from documents on the first conflict to those of the other without access to this specific cluster. We could say that this cluster has discovered an unexpected, and therefore informative facet in the data that would have been difficult for a human (without apriori knowledge of the two conflicts) to find. This intuitive explanation has been formally captured by the information theoretic scores that FICS has given to facets in this cluster.

We should also note that finding such hidden and informative associations quickly is important to homeland security applications.

(a) Expt. 1 Result C_{inf}

Term	norm(·)	F_C
israel	10.2	3
isra	7.6	3
armenian	8.1	2
arab	5.2	3
jew	4.3	1, 3
muslim	3.1	1, 2
turkish	2.3	2
palestinian	1.5	3
jewish	1.2	1,3
kill	1.2	1,2,3

(b) Expt. 2 Result C_{uninf}

Term	norm(·)	F_C
game	9.6	1,2
team	7.6	1,2
player	4.8	1,2
play	2.9	1,2
baseball	2.3	1
hockey	2.2	2
win	2.1	1,2
season	1.8	1,2
fan	1.6	1,2
score	1.5	1,2

Table 3. Table of terms, along with their L_1 norms and facets in which they occur for the result clusters of Expt. 1 and Expt. 2. Let us compare the two clusters. In C_{inf}, there are three frequent itemsets, and two facets. The terms that connect facets are (a) few, and (b) occur infrequently (i.e., have low L_1 norm). For example, the only term that connects both facets (i.e., all three frequent itemsets) "kill" which has the lowest L_1 norm. This makes facet connections unlikely, and the information content in these facets high. In sharp contrast, in C_{uninf}, there is one facet, and nearly all terms occur in its connection. This includes highly frequent terms such as "game", "team", etc. Intuitively, going from one frequent itemset in this facet to the other is easy due to several co-frequent terms between them.

8.2 Experiment 2: Use FICS to Find a Cluster Having Expected Concept Associations

Objective To find a cluster having few facets whose concept associations are expected. Some applications may require such "highly coherent" clusters.

Methodology In order to contrast with Expt. 1, we again use the dataset N20 for this experiment, with the same parameters.

Result The bottom two positions in both ranked lists $L_{\mathsf{FCI}}(\mathcal{C})$ and $L_{\mathsf{ENT}}(\mathcal{C})$ are taken by the same two facets. However, the cluster that has the bottom ranked facet also has two other facets that are ranked high, whereas the cluster that has the facet ranked second from the bottom in both lists has no other facets. Therefore, it is this latter cluster that has the least unexpected facets. We will call it C_{uninf}. The cluster summary for C_{uninf} is provided in Fig. 3 and Table 3(b).

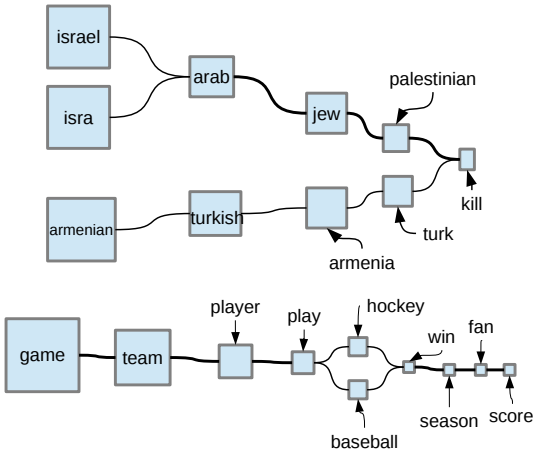


Fig. 3. C_{inf} (top, simplified for illustration) and C_{uninf} (bottom) of Expt. 1 and 2. C_{inf} connects two distinct historical themes separated by a significant period of time, and yet having geographical and structural similarities, resulting in informative facets. In contrast, C_{uninf} is "highly coherent:" it has a single facet, with a minor difference in the use of two sports in the two frequent itemsets that form that facet. The facet is easy to discover due to the high frequency of connecting terms in it.

Discussion We can interpret the expectedness of concept associations of cluster C_{uninf} as follows. There are two concepts in this cluster: the playing seasons of two sports (baseball and hockey), as viewed through their fans. The difference between baseball and hockey as it pertains to the discussions in C_{uninf} is minor. For example, there are several sentences in several documents in the cluster where both baseball and hockey fans are discussed in common. An example is the sentence "baseball fans can watch baseball and hockey fans can watch hockey." The single facet that links both the concepts of this cluster has several connecting terms, several of whom occur frequently in the cluster. Put another way, had we relaxed the threshold for frequent itemset slightly, both the concepts in the cluster would have merged into a single larger concept. This makes the cluster "highly coherent" conceptually.

In summary, the concepts represented by the two frequent itemsets of C_{uninf} are commonly present *together* in several documents, making their association neither very surprising nor hard to find. A cursory inspection of a few documents on either concept is likely to reveal this association.

8.3 Experiment 3: Use FICS to Find an Informative Multi-faceted Cluster

Objective To find an informative cluster that exhibits many disparate facets.

Methodology We use the dataset `REU` for this experiment. We cluster into $K=20$ clusters for a manageable list of clusters. We then identify clusters that have at least 5 facets. We ranked such clusters using $\mathsf{FCI}(\cdot)$ and $\mathsf{ENT}(\cdot)$.

Result For the purposes of illustration, we picked the most informative cluster having 5 facets. The cluster $C_{multi,inf}$ is identified as an informative multi-faceted cluster; it is illustrated in Fig. 4.

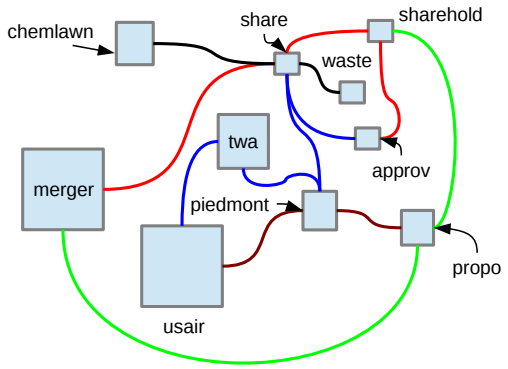


Fig. 4. Expt. 3: An informative multi-faceted cluster. The five prominent concepts in the cluster are indicated by five different colors. They depicts several corporate mergers/takeover attempts, some of whom happened simultaneously. The six facets resulting from concept connections have high information theoretic scores, meaning that they are unexpected (relative to the baseline statistics of the corpus), and therefore tedious to find by manual human inspection. Since the FICS framework also counts the number of facets, it can be used to focus on multi-faceted clusters. Emerging applications of data mining such as homeland security may require such features.

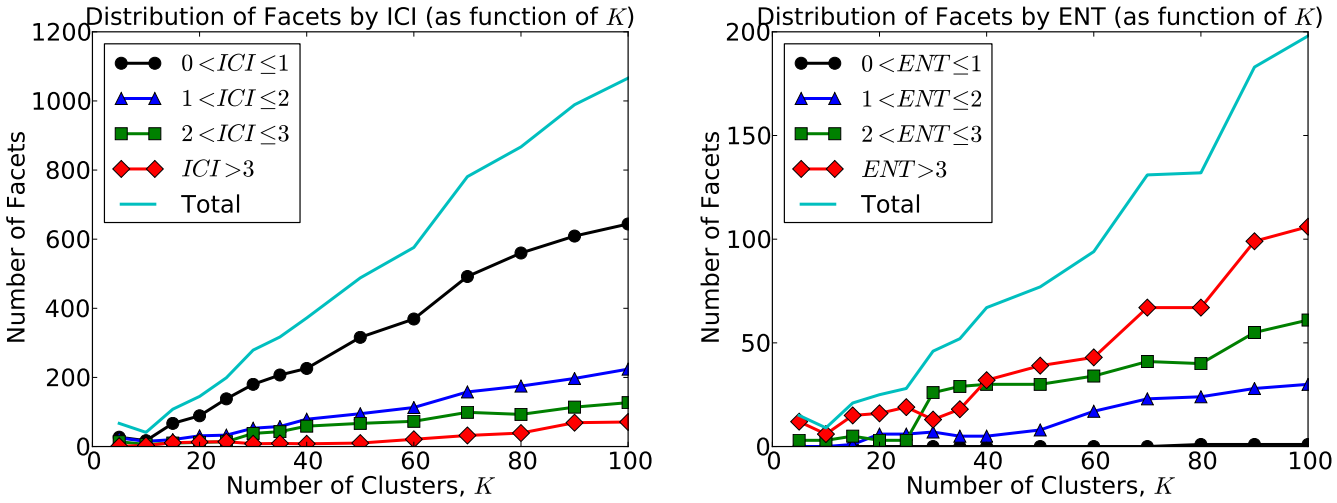
Discussion $C_{multi,inf}$ has 5 facets, each of them informative, having high $\mathsf{FCI}(\cdot)$ and $\mathsf{ENT}(\cdot)$. The facets are illustrated in the Fig. 4 as different colored edges. $C_{multi,inf}$ speaks of several merger and takeover attempts, some happening simultaneously. For example, the proposed TWA takeover of USAir at the same time that USAir was trying to acquire Piedmont.[5]

8.4 Experiment 4: Use FICS to find Informative Connections Across Different Slices of a Dataset

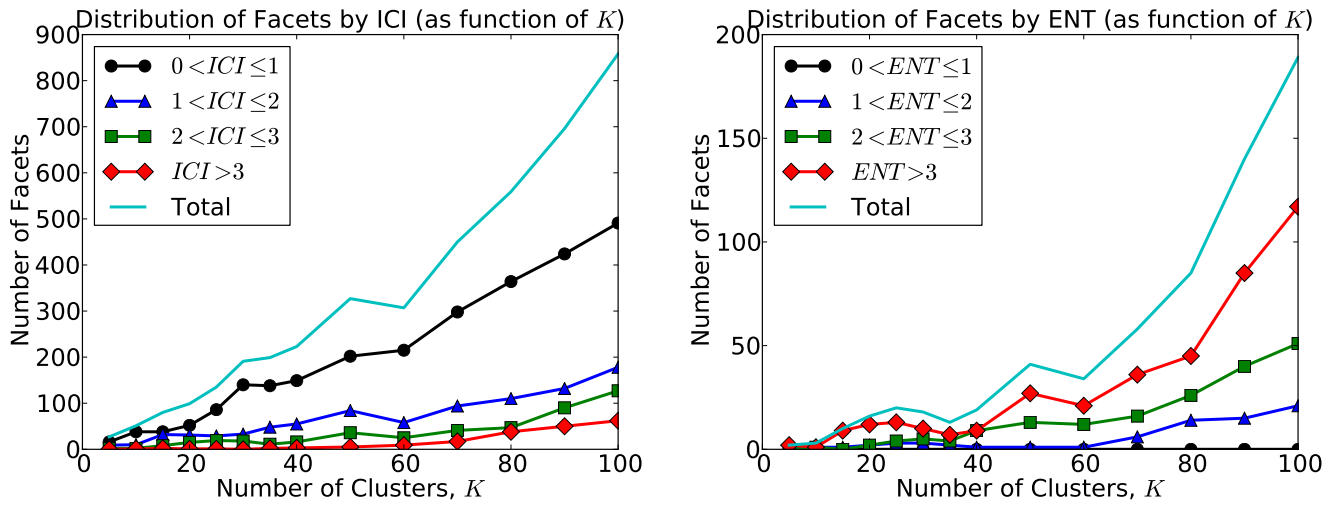
The components of the FICS framework can be used for non-standard data mining tasks, such as connecting different datasets, or different slices of a dataset.

Objective To find clusters that link one set of topics to different sets of topics in two different cross-sections of a dataset, and compare their informativeness.

[5] see http://tinyurl.com/b6b5ngn for an article speaking about this.



(a) N20



(b) REU R8

Fig. 5. Distribution of FCI(·) and ENT(·) for N20 and REU R8 datasets. We find a consistent pattern across all values of K: higher values of FCI(·) are rare. This indirectly corroborates our general assertion that these rare and highly informative facets should be identified early during processing of clusters. The distribution of ENT(·) concentrates at higher values, showing that high entropy in the weight distribution of facet components is relatively common. Thus, it is FCI(·) that is the dominant source of rare and informative facets.

Methodology We will use the R8 and R52 Train subsets of REU (see Table 8), both clustered into $K = 10$. We will identify an informative cluster in one dataset that also has a corresponding informative cluster in the other, such that both clusters have one set of topics in common (while other sets will not be common). In this way, we discover informative connections of this common set of topics with two different topics in two different datasets. In order to find a correspondence between clusters having a set of topics in common, we will look at overlaps of concepts across these clusters.

Result We find the topic of officials' remarks on Japan's export surplus to have informative connections in both datasets. In R52, it is related to OPEC's exports, where as in R8, it is related to the US's grain exports. The two clusters are summarized in Table 4.

Discussion Let us consider the connection in R52. It relates two of the most important export led growth stories of the 1980's: Japan's with the OPEC. In both cases, it prominently presents views of officials who regulated and controlled these growth stories. In R8, Japan's exports are tied to the US's grain exports, once more with emphasis on the roles of officials in both. In R8, the FCI(·) is 3.2 (indicating a highly informative connection), whereas in R52, it is also high at 2.5.

In Experiments 1 through 4, we have tried to convey to the reader how the FICS framework can be used in diverse manners to demonstrate informative facets and connections in textual data, as well as to identify "coherent" facets (as in Expt. 2). Our next two experiments focus, instead, on the distribution of these information theoretic scores across a dataset.

8.5 Experiment 5: Distribution of FCI(·) and ENT(·)

Objective Understand how FCI(·) and ENT(·) of connections is distributed, and how this distribution varies as we increase K.

(a) R8

Term	norm(·)	Facets
trade	21.8	2
japan	13.7	2
export	2.7	2
billion	2.2	2
import	2.1	2
offici	1.9	1,2
surplu	1.9	2
deficit	1.8	2
tariff	1.1	2
grain	1.1	1

(b) R52

Term	norm(·)	Facets
oil	9.4	3
trade	9	2
japan	6.4	2
crude	2.3	3
price	2	1,3
minist	1.6	1,2,3
offici	1.6	1,2
barrel	1.5	3
opec	1.5	1,3
export	1.2	1,2,3

Table 4. Connecting a concept across two clusters in two datasets: `REU` R8 and `REU` R52 Train (Expt. 4). In both, Japan's exports are connected, but to two different topics. See intext for details.

Methodology We use the datasets `N20` and `REU` R8 for this study. We cluster each dataset into K clusters. For each cluster, we compute the frequent itemsets, and their connections. For each connection, we compute $\mathsf{FCI}(\cdot)$ and $\mathsf{ENT}(\cdot)$, and plot their distributions in 4 ranges—(0,1],(1,2],(2,3], and above 3.

Results Shown in Fig. 5.

Discussion We can see that the informativeness component coming from $\mathsf{FCI}(\cdot)$ tends to concentrate at lower values of $\mathsf{FCI}(\cdot)$. Namely, higher values of $\mathsf{FCI}(\cdot)$ are rare. In comparison, facets tend to be proportionate, meaning that $\mathsf{ENT}(\cdot)$ concentrates towards the higher end. This pattern is consistent across all values of K.

8.6 Experiment 6: Correlation between Similarity and Informativeness

Objective To understand the relation (if any) between similarity of a cluster and its informativeness measures $\mathsf{FCI}(\cdot)$ and $\mathsf{ENT}(\cdot)$.

Methodology We use the datasets `N20` and `REU` R8 for this study. We cluster each dataset into K clusters. For each cluster, we compute the frequent itemsets, and their connections. For each connection, we compute $\mathsf{FCI}(\cdot)$ and $\mathsf{ENT}(\cdot)$. This gives us two arrays of numbers whose lengths are the number of connections in the K-way clustering. Now we create another array of the same length, whose entry at row i is the ISim of the cluster into which the connection at row i in the first two arrays falls. Finally, we take the Pearson correlation coefficient of these arrays.

Results Shown in Table 5.

Dataset	$\mathsf{FCI}(\cdot)$ vs. ISim	$\mathsf{ENT}(\cdot)$ vs. ISim	$\mathsf{FCI}(\cdot)$ vs. $\mathsf{ENT}(\cdot)$
`N20`	-0.18	-0.44	0.57
`REU` R8	-0.21	-0.27	0.57

Table 5. Pairwise Pearson correlation coefficients between$\mathsf{FCI}(\cdot)$, $\mathsf{ENT}(\cdot)$, and ISim

Discussion $\mathsf{FCI}(\cdot)$ and $\mathsf{ENT}(\cdot)$ are not strongly correlated with ISim. Due to this low correlation, we cannot infer much about $\mathsf{FCI}(\cdot)$ from just the ISim.

Note on Expt. 1 and 2 Even though there is a connection across a disparate pair of topics in C_{inf}, and in contrast C_{uninf} is highly coherent, the ISim of C_{inf} is *higher* than that of C_{uninf}. This also gives us a specific example instantiating the result of this experiment. Namely, we cannot use low ISim as a proxy for informativeness. This also clarifies a subtle point: a highly informative cluster is not the same as a "loose, disparate cluster." A loose cluster will have low ISim, but that is not a property that our informative clusters have. Indeed, as our examples illustrate, our informative clusters may be "tight" in that they have high ISim.

9 Conclusions and Future Work

We have studied the data mining task of measuring the information contained in concept connections in document clusters. Our approach to this task is based on sound statistical principles. We develop a general framework whose components can be used in multiple ways. Our choice of experiments is intended to convey the diverse application scenarios for this framework. Keeping with actual application scenarios that provided the impetus for this work, we have ensured that our approach results in an interpretable informativeness score.

This task provides ample opportunities for future work.

1. An immediate branching of the main task is to develop *subjective* measures for informativeness of clusters.

2. Clustering has a long tradition in information retrieval (IR). Can IR systems use informativeness scores of concept associations in a meaningful manner? The natural emphasis in IR is on relevance. Can relevance be augmented with informativeness in a principled manner?
3. Last, but perhaps most important from an application point of view, is the question of what other end-user workflows could benefit from a notion of informative document clusters.

Scope and Limitations of Our Work The multitude of interestingness measures for association rules points to the fact that different applications may require different notions of what is interesting. We feel that the situation will be similar for notions of informativeness for clustering. It is not reasonable to suppose that any single purely objective informativeness measure can capture all the intricacies of the notion of objective informativeness. That is why we have narrowed our focus to a specific type of informativeness that can be measured from concept representations and their associations.

We also wish to contrast our goal with that of finding *interesting* clusters. Interestingness may be in the eyes of the beholder: a person who is not interested in the middle east might not find our example of Experiment 1 interesting. However, the cluster is informative, in an information-theoretic sense, due to its concept associations. This statement can be made in a purely objective statistical manner. Also, subjective measures must be garnered to work in tandem with objective ones. In light of the above, we envision our information-theoretic measures to be combined with other measures of content ranking, especially subjective measures, by data miners. This is the manner in which these measures are currently being sought to be integrated into a large scale market leading commercial enterprise information management product.

Bibliography

[1] Charu C. Aggarwal and Philip S. Yu. A new framework for itemset generation. In Alberto O. Mendelzon and Jan Paredaens, editors, *PODS*, pages 18–24. ACM Press, 1998.

[2] Rakesh Agrawal, Tomasz Imieliński, and Arun Swami. Mining association rules between sets of items in large databases. *SIGMOD Rec.*, 22(2):207–216, June 1993.

[3] Rakesh Agrawal and Ramakrishnan Srikant. Fast algorithms for mining association rules in large databases. In *Proceedings of the 20th International Conference on Very Large Data Bases*, VLDB '94, pages 487–499, San Francisco, CA, USA, 1994. Morgan Kaufmann.

[4] Tom Brijs, Koen Vanhoof, and Geert Wets. Defining interestingness for association rules. In *Int. Journal of Information Theories and Applications*, volume 10, pages 370–376, 2003.

[5] Sergey Brin, Rajeev Motwani, Jeffrey D. Ullman, and Shalom Tsur. Dynamic itemset counting and implication rules for market basket data. *SIGMOD Rec.*, 26(2):255–264, June 1997.

[6] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory (2nd ed.)*. Wiley, 2006.

[7] Vinay Deolalikar. What is the shape of a cluster?: Structural comparisons of document clusters. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, CIKM '14, pages 1927–1930, New York, NY, USA, 2014. ACM.

[8] Byron E. Dom. An information-theoretic external cluster-validity measure. In *Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence*, UAI'02, pages 137–145, San Francisco, CA, USA, 2002. Morgan Kaufmann.

[9] Benjamin CM Fung, Ke Wang, and Martin Ester. Hierarchical document clustering using frequent itemsets. *SDM*, 3:59–70, 2004.

[10] Gartner. Forecast: Enterprise e-discovery software, worldwide, 2012-2017. 2012.

[11] Liqiang Geng and Howard J. Hamilton. Interestingness measures for data mining: A survey. *ACM Computing Surveys*, 38(3), September 2006.

[12] Maria Halkidi, Yannis Batistakis, and Michalis Vazirgiannis. On clustering validation techniques. *Journal of Intelligent Information Systems*, 17:107–145, 2001.

[13] Robert James Hilderman and Howard John Hamilton. *Knowledge discovery and interestingness measures: A survey*. Department of Computer Science, University of Regina, 1999.

[14] IBM. *IBM Intelligent Miner User's Guide, Version 1, Release 1*. 1996.

[15] Kleanthis-Nikolaos Kontonasios, Eirini Spyropoulou, and Tijl De Bie. Knowledge discovery interestingness measures based on unexpectedness. *Wiley Int. Review on Data Mining and Knowledge Discovery*, 2(5):386–399, September 2012.

[16] Hassan H Malik and John R Kender. High quality, efficient hierarchical document clustering using closed interesting itemsets. In *Proceedings of the 6th International Conference on Data Mining, ICDM'06*, pages 991–996. IEEE, 2006.

[17] Ken McGarry. A survey of interestingness measures for knowledge discovery. *The Knowledge Engineering Review*, 20(01):39–61, 2005.

[18] Balaji Padmanabhan and Alexander Tuzhilin. A belief-driven method for discovering unexpected patterns. In Rakesh Agrawal, Paul E. Stolorz, and Gregory Piatetsky-Shapiro, editors, *KDD*, pages 94–100. AAAI Press, 1998.

[19] Balaji Padmanabhan and Alexander Tuzhilin. Unexpectedness as a measure of interestingness in knowledge discovery. *Decision Support Systems*, 27(3):303 – 318, 1999.

[20] Gregory Piatetsky-Shapiro. Discovery, analysis, and presentation of strong rules. In *Knowledge Discovery in Databases*, pages 229–248. AAAI/MIT Press, 1991.

[21] Monographs MG 1208 Rand. Where the money goes: Understanding litigant expenditures for producing electronic discovery. 2012.

[22] Jeffrey Seifert. Data mining and homeland security: An overview. *CRS Report for Congress*, 2007.

[23] Abraham Silberschatz and Alexander Tuzhilin. What makes patterns interesting in knowledge discovery systems. *IEEE Transactions on Knowledge and Data Engineering*, 8(6):970–974, 1996.

[24] Craig Silverstein, Sergey Brin, and Rajeev Motwani. Beyond market baskets: Generalizing association rules to dependence rules. *Data Min. Knowl. Disc.*, 2(1):39–68, 1998.

[25] Pang-Ning Tan, Vipin Kumar, and Jaideep Srivastava. Selecting the right interestingess measure for association patterns.

[26] Pang-Ning Tan, Vipin Kumar, and Jaideep Srivastava. Selecting the right objective measure for association analysis. *Information Systems*, 29(4):293–313, 2004.