VICTORIA UNIVERSITY OF WELLINGTON
Te Whare Wananga o te Upoko o te Ika a Maui

School of Mathematical and Computing Sciences
Computer Science

PO Box 600
Wellington
New Zealand

Tel: +64 4 463 5341, Fax: +64 4 463 5045
Email: Tech.Reports@mcs.vuw.ac.nz
http://www.mcs.vuw.ac.nz/research

Universal Evaluation Method for Web Clustering Results

Daniel Crabtree, Xiaoying Gao, Peter Andreae

Technical Report CS-TR-05/3
September 2005

Abstract

Finding a set of web pages relevant to a user's information goal is difficult due to the enormous size of the internet. Search engines are able to find a set of pages that match the user's query, but refining the results of the search is still difficult and time consuming. Web clustering addresses this problem by presenting the user with clusters of related pages as refinement options. Many clustering algorithms have been developed and researchers need to be able to compare their effectiveness. The lack of a fair universal evaluation method has led to incomparable research and results. This paper identifies the requirements for evaluating the clusters produced by a web clustering algorithm and proposes a new method for a fair universal evaluation of clusterings to meet the requirements. The paper also shows how the new method can evaluate clusterings with diverse characteristics that are not directly comparable by previous methods.

Keywords: web clustering, evaluation

1 Introduction

The problem facing a user searching the web is the enormous size of the internet and the difficulty of identifying a small set of relevant web pages. Current search engines allow a user to retrieve pages that match a search query, but the number of results returned by a search engine is often huge, and many of the results may be irrelevant to the user's goal. Search engines attempt to order the results to present pages that are more useful earlier, but the user will generally need to refine their search by adding to or changing the query to filter out the irrelevant results. The large ordered list of results provides little assistance to the user in this difficult query refinement process — the user may need to retrieve and scan many of the pages to determine the topics of irrelevant pages that need to be excluded by the refined query.

A promising technique to address this problem is to organize the result set into clusters of semantically related pages so that the user can quickly overview the entire result set, and can use the clusters themselves to filter the results or refine the query. There are different kinds of possible clusterings of a result set, each has a set of characteristics: the clustering granularity may be coarse, so that there are just a few large clusters covering very broad topics, or fine, so that there are many small clusters of very focused topics; the clusters may be disjoint and constitute a partition of the results, or the clusters may overlap, so that the same page may appear in several clusters; the clustering may be "flat" so that all clusters are at the same level, or the clustering may be hierarchical so that lower-level clusters are subclusters of higher level clusters. Many clustering algorithms have been developed (eg, K-means [8], Hierarchical Agglomerative Clustering [1], Link and Contents Clustering [12], Suffix Tree Clustering [13], etc.) and different algorithms produce clusterings with different characteristics.

A critical requirement in the development of techniques for clustering web search results is to be able to fairly evaluate and compare clusterings with different characteristics and hence different algorithms. Currently used evaluation methods fail to meet this requirement. This paper presents a new evaluation method that meets this requirement by generalizing the "gold-standard" approach to use a richer kind of ideal clustering and by developing new measures of cluster quality and topic coverage.

The next section outlines previous methods. Section 3 specifies requirements on a universal clustering evaluation method and discusses the problems with existing methods and their measurements. Section 4 describes our proposed method and justifies why the method meets the requirements. Section 5 compares our method against mutual information using synthetic examples. Section 6 concludes the research and provides direction for future work.

[3, 7, 4]

2 Previous Methods

2.1 Methodology

There are two broad methodologies for evaluating clusterings. Internal quality [9] is a model-based, unsupervised approach that evaluates a clustering only in terms of a function of the clusters themselves. Internal quality is often used when there is no information about the desired output that can be used in the evaluation. External quality [9] is a model-free, semi-supervised approach that evaluates a clustering using external information (ie, information not available to the clustering algorithm), such as an ideal clustering, or its effectiveness in some applications. Where there is such external information (eg, ideal clustering of web search results), external quality is more appropriate because it allows the evaluation to reflect performance relative to the desired output.

2.2 Approaches

There are three main approaches to evaluation using the external methodology: gold-standard, task-oriented, and user evaluation. The first two are considered by [10]. Gold-standard approaches manually construct an ideal clustering with each ideal cluster labeled with a topic, which is then compared against the actual clustering. Task-oriented approaches evaluate how well some predefined task is solved. A common task is search result ordering: which involves labeling pages as relevant or irrelevant to a topic, reordering the search results using the clusters, and then evaluating the reordered results [13]. User evaluation approaches involve directly studying the usefulness for users and often involve observation, log file analysis, and user studies similar to those carried out in the user evaluation of Grouper I and Grouper II [13].

Task-oriented methods such as search result reordering have a bias towards the selected task. Task-oriented methods are by definition, biased towards some task and hence biased towards certain clustering characteristics, making them poor candidates for a fair universal clustering evaluation method. For example, search result reordering [13], has a bias towards small clusters, which tend to have higher quality. Randomly generating a perfect cluster of five pages is much more likely than generating a perfect cluster of fifty pages. In the extreme case of one cluster per page (singleton clustering), the clustering is evaluated as perfect, when clearly it is not.

User evaluation methods are not reproducible as they are reliant on the users, and even the same users are unlikely to produce identical results all the time. The results are often inconclusive and frequently have multiple valid interpretations. The selection of users introduces significant user bias, although this can almost be reduced to zero by using large samples, it then becomes prohibitively costly. The lack of reproducibility, large cost, and time involved in conducting user evaluations makes them poor candidates for a fair universal clustering evaluation method.

The most appropriate method is to use the external information about an ideal clustering to define a gold-standard. Given an ideal clustering, the problem is how to measure a clustering against it.

2.3 Measurements

Clusterings have been evaluated using a wide variety of measurements. This section outlines those most commonly used; their limitations will be discussed in the next section.

The rest of the paper uses the following notation: C is a set of clusters, T is a set of topics (the clusters of the ideal clustering), and D is a set of pages. c, t, and d are individual elements of C, T, and D respectively. D_c is the pages in cluster c, D_t is the pages in topic t, and $D_{c,t}$ is the pages in cluster c of topic t.

Purity [9] and F [8, 9, 2] are two clustering evaluation methods that are based on three standard information retrieval [11] measures: precision, recall, and f-measure.

$P(c,t) = \text{Precision} = \frac{|D_{c,t}|}{|D_c|}$

$R(c,t) = \text{Recall} = \frac{|D_{c,t}|}{|D_t|}$

$F(c,t) = \text{F-measure} = \frac{2*P(c,t)*R(c,t)}{P(c,t)+R(c,t)}$

Purity assumes that a cluster represents the topic with the highest precision. F assumes that a cluster represents the topic with the highest f-measure.

$Purity = \sum_{c \in C} \frac{|D_c|}{|D|} \max_{t \in T}\{P(c,t)\}$

$F = \sum_{c \in C} \frac{|D_c|}{|D|} \max_{t \in T}\{F(c,t)\}$

Entropy and Mutual Information (MI) are well founded in information theory [5]. Entropy[1] can be considered an advanced version of purity, which evaluates a single cluster by considering its distribution among all topics, rather than just one topic. MI[2] evaluates how closely the entire clustering

[1] In contrast to all other measurements considered, lower entropy is better. 0 is best, 1 is worst.

[2] There are also other forms of MI for a clustering.

matches the ideal clustering.
$Entropy(c) = -\sum_{t \in T} P(c,t) log_{|T|} P(c,t)$
$MI = \frac{2}{|D|} \sum_{c \in C} \sum_{t \in T} |D_{c,t}| log_{|C||T|} (\frac{|D_{c,t}||D|}{|D_c||D_t|})$

Some methods [8] use entropy weighted by cluster size for overall evaluation, while others [9] use MI for overall evaluation. Another information theoretic measurement for comparing clusterings is variation of information (VI) [6], which combines MI and Entropy to define a metric on clusterings.

3 Requirements and Limitations

This section introduces the proposed requirements of a fair universal clustering evaluation method, and then identifies some of the limitations with previous methods.

3.1 Evaluation Requirements

A clustering is perfect if it is identical to the ideal clustering. A clustering can fail to be perfect in two ways, (1) some clusters may be of poor quality by way of not exactly matching ideal clusters (topics) and (2) the clustering may not include (cover) all the pages in the ideal clustering. There is often a trade off between these two attributes — algorithms can often be tuned to perform one well, at the cost of the other. Different users, and different applications, will each put different weight on each attribute depending on their needs. For instance, typical web users may place equal weight on both; cell-phone users may want higher quality, but accept lower coverage; researchers may want higher coverage, but accept lower quality. Therefore, good evaluation methods must measure each attribute separately (the measurements can be combined later to give a single application specific measurement). Although not considered in this research, other factors (eg, computational complexity, run time, memory requirements, etc.) should also be measured.

When considering the fairness of a clustering evaluation method, it is useful to keep two criteria in mind. (1) An experimenters freedom to construct ideal clusterings that favor particular algorithms should be minimized. (2) The evaluation method and measurements should not inherently favor any particular algorithm. One way to ensure (2) is to minimize the method bias towards clusterings with particular characteristics (cluster granularity: coarse or fine, clustering structure: hierarchical or flat, disjoint or overlapping, and cluster size: large or small): if the bias towards the different possible characteristics of a clustering is minimized, then so is the bias towards the algorithms that produce those clusterings.

3.2 Limitations of Previous Methods

Previous methods use an overly constrained ideal clustering structure for their gold-standard that allows only one ideal clustering. As clusterings can have different characteristics, there are many equally ideal clusterings that an experimenter could pick, giving the experimenter a lot of freedom to favor a particular method. Further, clusterings with characteristics other than those of the chosen ideal clustering are penalized, creating method bias towards certain clusterings.

Overall measurements used by the current methods are not satisfactory. There is no well-defined independent measure of coverage. Entropy and purity measure quality, have method bias towards small clusters, and are maximized by a set of singleton clusters. Entropy[3], F, and purity are related to the proportion of pages in the largest topic, which can lead to very high performance for useless clusterings (eg, a single cluster containing all pages), a form of method bias. Entropy, F, and purity measures are weighted by cluster size, so if a great disparity between the large and small clusters exists, these measures provide little indication of the performance of smaller clusters, creating method

[3] When topics have non-uniform size

bias towards the performance of large clusters. MI, VI, and F wrap up quality and coverage with a single measure, making it impossible to interpret the two attributes separately. MI and VI require that clusters and topics partition the pages, so clusters and topics may not overlap, and entropy requires this of topics, creating method bias.

4 New Method - QC4

QC4 (Quality, Coverage, and 4 Overall Measurements) addresses the problem of an overly constrained ideal clustering by introducing a new, more general ideal clustering that describes all ideal clusterings and new overall measurements are developed that fairly characterize all clusterings in terms of quality (1) and coverage (2). The new overall measurements meet the requirements by avoiding bias towards clusterings with particular characteristics.

4.1 A General Ideal Clustering

A general ideal clustering is a hierarchy of idealized clusters, termed topics, that is created by a human expert based on the pages to be clustered. Pages often have many topics that naturally overlap and that form hierarchies with different degrees of topic granularity. Because of this, the hierarchies created should often include topics at multiple levels and the topics should often overlap to some extent. Every possible ideal clustering of the pages should be a subset of the hierarchy. Topics at the top of the hierarchy cover very broad topics, and topics lower in the hierarchy, have progressively finer granularity. Each subtopic is a subset of a single parent. Topics have a level defined by their depth in the hierarchy; the topics with no parents are the top-level of the hierarchy (level 1). Topics at any given level may overlap.

To ensure clusterings with clusters of different granularity are evaluated fairly, all pages must be assigned topics at all levels. If a subtopic exists, pages in its parent but not in that subtopic must be in some other subtopic of that parent. To ensure topics have sub-topics at all levels, topics without sub-topics are duplicated at lower-levels as children of themselves. It is very unusual to have only one topic at the top-level, if there is and there are lower-levels, remove the top-level. It is also unusual to have particularly small topics, as topics get smaller, the impact on the evaluation accuracy diminishes and the probability of mistakes in the topics increase.

All pages should have a topic and all topics should be non-empty. Searches often find some pages that are very distinct and which are often completely unrelated or erroneous (ie, really should not have been in the result set). These often do not share any sensible topic with more than a few pages; these pages are termed outliers and are placed in a special topic, termed the outlier topic. The outlier topic is always disjoint from all other topics and is identical at all levels. Typically, the number of pages in the outlier topic is small.

As the general ideal clustering defines all ideal clusterings, a lot of the experimenter freedom granted by previous methods for selecting a single constrained ideal clustering and the method bias towards the specific selected clustering is eliminated. The only problem now is to define measurements that fairly characterize clusterings according to the general ideal clustering.

4.2 Characterize Clusterings

QC4 assumes that a clustering algorithm generates a set of clusters (duplicate clusters can be trivially removed). Different applications handle outliers in different ways. To avoid method bias towards the handling of outliers, clusters predominantly containing pages from the outlier topic are excluded from the cluster set, the topic that represents them is not considered for coverage, and they negatively affect the quality of any remaining clusters of which they are members.

In addition to the notation in section 2.3, the rest of the paper uses the following notation: L is the set of levels from the topic hierarchy (eg, 1, 2, 3) and l is an individual element of L. C_d is the clusters containing page d, T_l is the topics at level l, T_d is the topics containing page d, and $T_\emptyset$ is a set containing the outlier topic. $sub(t)$ is the set of all descendants of topic t. $lvl(t)$ is the lowest level of topic t.

4.2.1 Basic Measurements

Precision $P(c,t)$, recall $R(c,t)$, and f-measure $F(c,t)$ are as defined earlier. Precision provides a measure of the probability that a cluster describes a given topic. F-measure provides a measure of similarity between a cluster and a topic. The topic with the highest f-measure for a cluster is thus the single most similar topic. Entropy compares a cluster against all topics and measures how close a cluster is to a single topic by measuring how much information would be required to refine the cluster into the separate topics it represents.

4.2.2 Cluster Quality

Cluster Quality is a measure of how closely a cluster matches a single topic. Cluster Quality, QU, is measured using a modified entropy measurement, E, scaled by a measure of the information the cluster provides about a single topic, I.

$QU(c) = I(c)E(c)$

The problem is that entropy does not work with overlapping topics due to over counting. Two kinds of overlap need handling: overlap between topics at different levels, and overlap between topics at the same level.

The overlap between levels is handled by evaluating measurements across only topics from one level, the level that contains topics that are the most similar to the cluster. As topics within a level typically have similar granularity, the level with the most similar topics is likely to be the level with the topic that is most similar to the cluster. So the lowest level of the topic with maximum f-measure is used as the level of a cluster.

$L(c) =$ cluster-level[4]$= lvl(argmax_{t\in T\setminus T_\emptyset}\{F(c,t)\})$

$E(c) = \max_{t\in T_{L(c)}}(1+E'(c,t))$

$E'(c,t_m) = \sum_{t\in T_{L(c)}} P'(c,t,t_m)log_{|T_{L(c)}|}P'(c,t,t_m)$

The overlap between topics at the same level is handled by using modified precision, P'. Due to over counting, the sum of precision can be greater than one with overlapping topics, causing perfect clusters to receive sub-perfect entropy. So precision is modified to cope with overlapping topics, while preserving the property that an ideal cluster gets perfect entropy. Pages in the best topic are treated as being in only that topic, while topics containing the remaining pages that would normally be over counted are normalized to counteract the over counting. The best topic, t_m, is the topic that maximizes the resulting entropy, and a nice property of this is that a topic that overlaps with many topics is preferred to a completely disjoint topic of the same precision.

$P'(c,t,t_m) =$

$$\begin{cases} \frac{|D_{c,t}|}{|D_c|} & if\{t = t_m\} \\ \frac{(|D_c|-|D_{c,t_m}|)|D_{c,t}\setminus D_{c,t_m}|}{|D_c|\sum_{t'\in T_{L(c)}\setminus\{t_m\}}|D_{c,t'}\setminus D_{c,t_m}|} & otherwise \end{cases}$$

Using the modified entropy, cluster quality measures how close a cluster is to a single topic and allows clusters of different levels of granularity, and disjoint and overlapping topics to be handled fairly. However, entropy does not consider all cluster quality information. Quality should only be perfect when a cluster exactly matches a single topic. A single cluster containing pages from one

[4] If multiple topics maximize F, the one with lowest level is selected.

topic is preferable to two or more smaller clusters containing the same pages from that topic. But perfect entropy is given to any cluster that contains purely pages from one topic. In addition, clusters that are similar to a random selection of pages from the result set provide almost no information beyond the un-clustered result set, and should receive near zero quality. However, entropy is related to the proportion of pages in the largest topic in randomly constructed clusters. To deal with these two cases, a new measure, I, is defined to scale entropy for incorrectly handled clusters. $I1$ deals with comparing cluster size and topic size, and $I2$ deals with randomly constructed clusters. I uses the minimum of the two, to avoid applying both to any one cluster.

$I(c) = \min\{1, I1(c), I2(c)\}$

The correlation between cluster size and topic size is related to the recall of the topic represented by the cluster. However, assuming the user is equally likely to desire any given topic, and will only consider a subset of the pages in any cluster, an extra percent recall of a large topic is less likely to benefit the user than an extra percent recall of a small topic. To address this, for any given recall, the recall is scaled upwards in relation to the log of the topic size. Therefore, ceteris paribus, the benefit of increasing recall by 1% in a large cluster is less than the benefit in a small cluster. To handle clusters containing pages from multiple topics, the scaled recall is weighted by the modified precision for each topic represented in the cluster. Using modified precision addresses the overlap between clusters at the same level; overlap between levels is addressed by considering only topics from the clusters level.

$I1(c) = \max_{t_m \in T_{L(c)}} \{\sum_{t \in T_{L(c)}} P'(c, t, t_m) I1'(c, t)\}$

$I1'(c, t) = 2^{\frac{R(c,t)-1}{R(c,t) log_2 |D_t|}}$

Randomly constructed clusters are identified using a modified version of MI. When considering disjoint topics, on average a random cluster will have a very similar fraction of pages from each topic. When the fractions are the same, MI is 0, when the fractions are very similar, MI is almost 0. Therefore, MI has the desired property, but MI fails when topics overlap due to over counting. The over counting is avoided by splitting the topics from the clusters level into a set of non-empty disjoint regions, REG, where every page is in exactly one region. This can be visualized as the regions of a Venn diagram of the topics. For example, if two topics overlap, there are three regions: the intersection of the two topics, and the two set differences.

$$REG(c) = \{r \subseteq D | (\exists T_\alpha \subseteq T_{L(c)} (|r| > 0 \wedge r = \\ \bigcap_{r' \in T_\alpha} D_{r'} - \bigcup_{r'' \in T_{L(c)} \setminus T_\alpha} D_{r''})\}$$

Another problem with MI is that the maximum value varies between 0 and 1 depending on the topic size distribution. To solve this, $I2$, a variation of MI is defined, that normalizes the modified MI by 5% of the minimum modified MI of an ideal cluster (topic) from the level of the cluster being evaluated. This ensures consistent results and the 5% threshold ensures that only clusters that are very close to random or that are very small are scaled down. This is acceptable, as $I2$ should scale random clusters, and although very small clusters are already handled by $I1$, there is no harm in handling them again with $I2$; $I1$ is usually less than $I2$ for very small clusters. The normalization and consideration of just a single cluster allows some scaling terms and the sum over clusters to be eliminated from the traditional MI, to produce $I2'$ where D' is either D_c or D_t. Note: If $|REG| \leq 1$ then $I2(c)$ is 1.

$$I2(c) = \frac{I2'(D_c, c)}{0.05 \min_{t \in T_{L(c)} \setminus T_\emptyset} \{I2'(D_t, c)\}}$$

$$I2'(D', c) = \sum_{r \in REG(c)} |D' \cap r| log_{|REG(c)|} \frac{|D' \cap r||D|}{|D'||r|}$$

4.2.3 Topic Coverage

Topic Coverage is a measure of how well the pages in a topic are covered. Topic Coverage, CV, is measured by the fraction of pages from the topic that are present in some cluster, where each pages contribution is weighted to reflect how well that page is covered.

A page is covered to some extent in a topic if there is a cluster that contains the page. A page is covered to the extent that the best subset of clusters containing the page describe the topic, an appropriate set of that topics children, or an appropriate set of that topics descendents by recursion. Page coverage is thus defined recursively by PC as the maximum of: the precision of the cluster that best describes the topic and contains that page, or the average page coverage in the children of the topic that contain the page.

$CV(t) = \frac{\sum_{d \in D_t} PC(d,t,1)}{|D_t|}$

$PC(d,t,l) = \frac{\sum_{t' \in T_l \cap T_d \cap sub(t)} \max\{PC'(d,t',l), PC(d,t',l+1)\}}{|T_l \cap T_d \cap sub(t)|}$

$PC'(d,t,l) = \max_{c \in \{c \in C_d | L(c)=l\}} \{P(c,t)\}$

Initially this definition of topic coverage may seem unusual, this is because it does not make much sense to talk about the coverage of some lower-level topic; when computed the overall coverage of a clustering, only the top-level coverages are used, which inherently considers the lower-level topic coverage by way of the recursive process. Any clustering that is an ideal subset of the topic hierarchy is given perfect coverage, any negative alteration to a clustering decreased coverage by an amount proportionate to the amount of degradation (eg, a page that is not present in any cluster has greater degradation on coverage than a page that is in a cluster that poorly describes the topic containing the page), and overlapping topics that are partially covered are appropriately penalized by averaging the individual coverage of each.

4.2.4 Overall Measurements

Overall measurements measure the cluster quality and topic coverage across the entire clustering. Individual cluster quality is correctly reflected by QU, and individual top-level topic coverage is correctly reflected by CV. The methods both fairly evaluate disjoint and overlapping topics, and topics of varying granularity without bias. Hierarchical and flat clusterings are considered fairly, as hierarchical clusterings can be flattened into a flat clustering and the clusters will be evaluated fairly as clusters of varying granularity are treated without bias. Method bias is thus minimized towards clusterings with any of these characteristics.

There are four overall measurements that characterize a clustering — two for quality and two for coverage. The overall quality measures consider the quality across all clusters, while the overall coverage measures consider the coverage across all topics, as top-level topic coverage reflects the coverage in lower-level topics.

Since topic sizes and thus desired cluster sizes can vary dramatically, cluster quality and topic coverage are combined in two ways to reflect the characteristics of clusterings with clusters of different sizes, minimizing method bias towards cluster size. Average measures place equal weight on every cluster and topic, weighted measures weight clusters and topics by their size, giving more emphasis to large clusters. In evaluating an web clustering algorithm for a particular application, a single appropriately weighted combination of the four overall measurements should be used.

$AQ =$ average quality $= \frac{\sum_{c \in C} QU(c)}{|C|}$

$WQ =$ weighted quality $= \frac{\sum_{c \in C} QU(c)|D_c|}{\sum_{c \in C} |D_c|}$

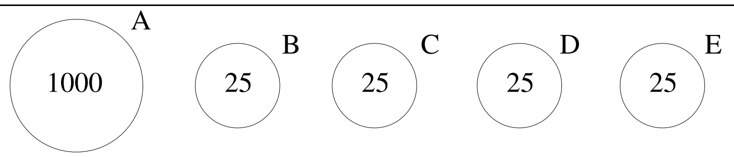
$AC =$ average coverage $= \frac{\sum_{t \in T_1 \setminus T_\emptyset} CV(t)}{|T_1 \setminus T_\emptyset|}$

$WC =$ weighted coverage $= \frac{\sum_{t \in T_1 \setminus T_\emptyset} CV(t)|D_t|}{\sum_{t \in T_1 \setminus T_\emptyset} |D_t|}$

In summary, the required clustering attributes (quality and coverage) are measured independently. The bias towards particular clusterings in previous gold-standard methods has been avoided: by minimizing experimenter freedom and method bias by defining a new general ideal clustering that allows all ideal clusterings to be defined simultaneously, and by minimizing the method bias in the measurement of the attributes by minimizing the bias towards the different characteristics a clustering can have. Therefore, the new evaluation method proposed meets the evaluation requirements set out earlier.

5 Comparison between QC4 and MI

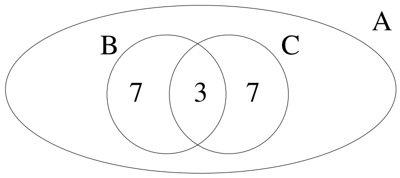
This section compares QC4 and MI using example clustering cases shown in Figures 1 and 2. The Venn diagrams represent the ideal clusterings by showing the topics labeled A-E, and the number of pages in each region. There are no outlier topics. In figure 2, A is a top-level topic, with two sub-topics, B and C. The columns show: case number, clustering description, the four QC4 measurements, and MI. All clusters are disjoint, except in cases (2), (11) and (12). The cluster description is explained using the following example: for ideal clustering shown in Figure 2, the clustering $(3B \cap C, 2C)$ 4x $(1B)$ would represent a cluster with 2 pages from topic C and 3 pages from the intersection of topics B and C, and 4 disjoint clusters each with 1 page from topic B.



		AQ	WQ	AC	WC	MI
(1)	(1000A)(25B)(25C)(25D)(25E)	1.000	1.000	1.000	1.000	0.268
(2)	(1000A)(25B)(25C)(25D)(25E)(25B,15C,15D)	0.887	0.968	1.000	1.000	0.333
(3)	(900A)	0.992	0.992	0.180	0.818	0.097
(4)	(900A)(5B)(5C)(5D)(5E)	0.639	0.983	0.340	0.836	0.091
(5)	(107A)(10B)(10C)(10D)(10E)	0.751	0.625	0.341	0.134	0.091
(6)	1000X(1A)25X(1B)25X(1C)25X(1D)25X(1E)	0.003	0.003	1.000	1.000	0.100
(7)	(1000A,25B,25C,25D,25E)	0.000	0.000	0.200	0.829	0.000
(8)	(170A,5B,5C,5D,5E)	0.034	0.034	0.035	0.139	0.068
(9)	(450A)(450A)	0.919	0.919	0.180	0.818	0.068

Figure 1: QC4 vs. MI

Case (1) shows a perfect clustering, all four QC4 measurements are correctly 1. But the MI is less than one, since the maximum MI depends on the page distribution among topics. This is unsatisfactory as there is no way to know how good a clustering is without knowing the maximum MI as a basis for comparison. Case (2) adds a low quality cluster to the perfect clustering (1), however, MI mistakenly evaluates (2) to be better than (1), this is due to the overlap between the clusters which MI does not handle. Case (2) also shows a key difference between QC4 quality and coverage measures. The addition of a low quality cluster negatively affects quality, but does not affect coverage; this is because coverage only considers the best clusters for each page and so the low quality cluster is ignored by coverage. Comparing (3) to (4), QC4 correctly shows the characteristics as an increase in coverage and a drop in quality, however, MI suffers in (4), the superior clustering. MI shows no difference between (4) and (5), however, there is a significant difference in the size of the clusters and this is shown by QC4, the weighted coverage has dropped significantly, due to the drop in coverage of the large cluster, highlighting the importance of having both average and weighted measures. Case (6) shows a singleton clustering, this correctly has almost 0 quality from QC4. However, MI does not penalize the singleton clustering enough and mistakenly shows it to outperform the reasonable clusterings (3), (4), and (5). Case (7) has all pages in one cluster, and (8) has almost the same fraction from each topic, which is close to a random cluster, both are bad clusterings with no information beyond the original result set. Both QC4 and MI correctly reflects these as 0 and near 0. Comparing (9) to (3) shows that splitting a cluster into two small clusters correctly decreases QC4 quality and MI.



		AQ	WQ	AC	WC	MI
(10)	(7B, 3B⋂C, 7C)	1.000	1.000	1.000	1.000	0.000
(11)	(7B, 3B⋂C)(3B⋂C, 7C)	1.000	1.000	1.000	1.000	0.432
(12)	(7B, 3B⋂C, 7C)(3B⋂C,7C)	1.000	1.000	1.000	1.000	0.216
(13)	(7B)(3B⋂C, 7C)	0.957	0.965	0.938	0.938	0.460
(14)	(2B, 3B⋂C, 7C)	0.317	0.317	0.502	0.502	0.188

Figure 2: QC4 vs. MI

MI cannot handle overlapping topics or clusters and as such it fails miserably with the cases in Figure 2. Cases (10), (11), and (12) each show one perfect clustering, and each is correctly evaluated as perfect by QC4. But MI gives a different answer for each, and in fact, MI gives 0 to one of the perfect clusterings, while it gives the best evaluation to a non perfect clustering (13). (10) and (11) show that QC4 handles the perfect clusterings at different levels correctly, while (12) shows clusters from multiple levels are handled correctly by QC4, a situation that arises frequently in hierarchical clusterings. Cases (13) and (14) show clusters that are evaluated against lower-level topics. (13) is correctly penalized in coverage for failing to cover the overlap in both sub-topics, while its quality is also penalized as one of its clusters is not identical to a topic. (14) is correctly penalized in coverage for covering only half of pages in the topics, while its quality is penalized for not purely representing a single topic.

This comparison shows that the QC4 method significantly outperforms MI across a variety of different synthetic conditions that simulate many of the different clustering characteristics.

6 Conclusions

This paper introduced QC4, a new evaluation method and justified why it is a fair universal clustering evaluation method. QC4 minimizes experimenter freedom and method bias by generalizing the gold-standard approach to use a more general ideal clustering that describes all ideal clusterings. QC4 introduces four new overall measurements that can universally characterize clusterings with different characteristics (cluster granularity: coarse or fine, clustering structure: hierarchical or flat, disjoint or overlapping, and cluster size: large or small) fairly in terms of cluster quality and topic coverage. It is also shown that QC4 significantly outperforms MI on many synthetic test cases that cover a broad range of clustering characteristics.

In the future, performance measurements such as computational complexity, run time, memory requirements, etc., need consideration. Standard test data and benchmark QC4 results for existing clustering algorithms also need to be developed.

References

[1] R. Ali, U. Ghani, and A. Saeed. Data clustering and its applications. `http://members.tripod.com/asim_saeed/paper.htm`.

[2] W. chiu Wong and A. Fu. Incremental document clustering for web page classification. In *IEEE 2000 Int. Conf. on Info. Society in the 21st century: emerging technologies and new challenges (IS2000), Japan*, November 2000.

[3] M. Halkidi, Y. Batistakis, and M. Vazirgiannis. On clustering validation techniques. *Journal of Intelligent Information Systems*, 17(2-3):107–145, December 2001.

[4] R. Koschke and T. Eisenbarth. A framework for experimental evaluation of clustering techniques. In *Proceedings of the 8th International Workshop on Program Comprehension*, page 201, 2000.

[5] D. J. Mackay. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, 2003.

[6] M. Meila. Comparing clusterings. Technical Report 418, Department of Statistics, University of Washington, 2002.

[7] B. S. Mitchell and S. Mancoridis. Craft: A framework for evaluating software clustering results in the absence of benchmark decompositions. In *Working Conference on Reverse Engineering*, pages 93–102, 2001.

[8] M. Steinbach, G. Karypis, and V. Kumar. A comparison of document clustering techniques. In *KDD Workshop on Text Mining*, 2000.

[9] A. Strehl. *Relationship-based Clustering and Cluster Ensembles for High-dimensional Data Mining*. PhD thesis, Faculty of the Graduate School of The University of Texas at Austin, 2002.

[10] P. Tonella, F. Ricca, E. Pianta, C. Girardi, ITC-irst, G. D. Lucca, A. R. Fasolino, P. Tramontana, U. di Napoli Federico II, Napoli, and Italy. Evaluation methods for web application clustering. In *5th International Workshop on Web Site Evolution, Amsterdam, The Netherlands*, September 2003.

[11] C. J. van Rijsbergen. *Information Retrieval*. Butterworths, London, 1979.

[12] Y. Wang and M. Kitsuregawa. Evaluating contents-link coupled web page clustering for web search results. In *Proceeding of 11th International conference on Information and Knowledge Management (CIKM 2002), McLean, VA, USA. ACM Press.*, pages 499–506, 2002.

[13] O. E. Zamir. *Clustering Web Documents: A Phrase-Based Method for Grouping Search Engine Results*. PhD thesis, University of Washington, 1999.