QC4 - A Clustering Evaluation Method

Daniel Crabtree, Peter Andreae, and Xiaoying Gao

School of Mathematics, Statistics and Computer Science
Victoria University of Wellington
New Zealand
daniel@danielcrabtree.com, pondy@mcs.vuw.ac.nz, xgao@mcs.vuw.ac.nz

Abstract. Many clustering algorithms have been developed and researchers need to be able to compare their effectiveness. For some clustering problems, like web page clustering, different algorithms produce clusterings with different characteristics: coarse vs fine granularity, disjoint vs overlapping, flat vs hierarchical. The lack of a clustering evaluation method that can evaluate clusterings with different characteristics has led to incomparable research and results. QC4 solves this by providing a new structure for defining general ideal clusterings and new measurements for evaluating clusterings with different characteristics with respect to a general ideal clustering. The paper describes QC4 and evaluates it within the web clustering domain by comparison to existing evaluation measurements on synthetic test cases and on real world web page clustering tasks. The synthetic test cases show that only QC4 can cope correctly with overlapping clusters, hierarchical clusterings, and all the difficult boundary cases. In the real world tasks, which represent simple clustering situations, QC4 is mostly consistent with the existing measurements and makes better conclusions in some cases.

1 Introduction

Comparing the performance of different clustering algorithms in some problem domains (i.e. web page clustering) has been problematic. Different algorithms produce clusterings with different 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 sub-clusters of higher level clusters. As a result, many of the existing evaluation methods are biased towards algorithms that produce clusterings with certain characteristics. An evaluation method that fairly evaluates clusterings with different characteristics is needed; so that all clustering algorithms can be compared with a consistent method.

An example clustering domain is web page clustering, which helps users find relevant web pages by organizing the search result set from a search engine into clusters of semantically related pages. These clusters provide the user with an

overview of the entire result set, and the clusters can be selected to filter the results or refine the query. Many clustering algorithms have been applied to web page clustering: K-means [1], Hierarchical Agglomerative Clustering [2], Link and Contents Clustering [3], Suffix Tree Clustering (STC) [4], Extended Suffix Tree Clustering (ESTC) [5], and Query Directed Clustering (QDC) [6]. A survey of clustering algorithms can be found in [7].

Many evaluation methods [1, 4, 7–11] are used to evaluate web clustering algorithms, but the results are often incomparable. There is probably no standard method because web page clustering algorithms produce clusterings that exhibit different characteristics, making web clustering an ideal application for an evaluation method that handles clusterings with different characteristics.

This paper proposes QC4, a new clustering evaluation method. Our preliminary research on QC4 was very briefly introduced in a short paper [12]. This paper further develops the full specifications of QC4, and evaluates it against existing measurements on synthetic test cases and real world web clustering tasks. QC4 allows clustering algorithms that produce clusterings with vastly different characteristics to be compared by generalizing the "gold-standard" approach to use a new structure for ideal clusterings and by developing new measures of quality and coverage. QC4 is currently targeted at web clustering, but is easily adapted to any domain where clusterings have different characteristics.

The next section discusses the related work. Section 3 describes and specifies QC4's richer ideal clustering structure and measurements. Section 4 evaluates QC4 by comparing it against the standard evaluation measurements using synthetic test cases and using nine clustering algorithms on four web page search result clustering tasks. Section 5 concludes the research and provides direction for future work.

2 Related Work

2.1 Approaches to Evaluation

There are two broad methodologies for evaluating clusterings. Internal quality [7, 8] evaluates a clustering only in terms of a function of the clusters themselves. External quality [7, 8] evaluates a clustering using external information, such as an ideal clustering. When external information is available, external quality is more appropriate because it allows the evaluation to reflect performance relative to the desired output.

There are three main approaches to evaluation using the external methodology: gold-standard [9], task-oriented [9], and user evaluation [4]. Gold-standard approaches manually construct an ideal clustering, which is then compared against the actual clustering. Task-oriented approaches evaluate how well some predefined task is solved. 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 [4].

Task-oriented methods have a bias towards the selected task. For example, search result reordering [4], which involves reordering the search results using the

clusters, 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 very difficult to reproduce as they are dependent on the users. The large cost, and time involved in conducting good user evaluations is also a significant drawback. The lack of reproducibility, large cost, and time involved in conducting user evaluations makes them poor candidates for a standardized clustering evaluation method.

Therefore our evaluation method uses external information in the form of an ideal clustering to define a gold-standard and measures a clustering against this ideal clustering.

2.2 Measurements

This section discusses the measurements most commonly used to evaluate a clustering against an ideal clustering in the web clustering domain. We refer to the clusters of the ideal clustering as topics, to distinguish them from the clusters of the clustering being evaluated.

A perfect clustering matches the ideal clustering. A clustering can be less than perfect in two ways: some clusters may be of poor *quality* because they do not match any topics well, and the clustering may not include (*cover*) all the pages in the ideal clustering. There is often a tradeoff between quality and coverage, and algorithms can often be tuned to achieve one well at the cost of the other. Good overall evaluation methods must measure both factors.

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 both cluster c and topic t. C_d is the set of clusters containing page d and C_t is the set of clusters that best match topic t: $C_t = c_i | \operatorname{argmax}_{t_j}(D_{c_i,t_j}) = t$.

Precision and recall are common measurements used in information retrieval [13] for evaluation. The precision, $P(c,t)$, of a cluster relative to a topic is the fraction of the pages in the cluster that are also in the topic. Whereas the recall, $R(c,t)$, is the fraction of the pages in the topic that are in the cluster. The F-measure [1, 8, 11] combines precision and recall with equal weight on each.

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

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

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

Purity is the precision of a cluster relative to its best matching topic. Because the pages in a topic may be included in several clusters, recall is seldom used for clustering. However, we could define the recall of a topic to be the total coverage of a topic among all clusters that best match that topic. F-measure is the f-measure of a cluster relative to its best matching topic.

$Purity(c) = \max_{t \in T}\{P(c,t)\}$
$Recall(t) = |\bigcup_{c \in C_t} D_{c,t}|/|D_t|$
$F(c) = \max_{t \in T}\{F(c,t)\}$

The Entropy and Mutual Information measures [1, 8] are based on information theory [14]. The Entropy measure is the average "narrowness" of the distribution of the pages of a cluster among the topics. More precisely, it is the amount of information required to refine the cluster into the separate topics it represents. Mutual Information (MI) is an average of a measure of correspondence between each possible cluster topic pair.

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

Average Precision (average purity over clusters), Weighted Precision (cluster size weighted average purity over clusters), Average Entropy (average over clusters), and Weighted Entropy (cluster size weighted average over clusters) [1] can be used for overall quality evaluation. Average Recall (average over topics) and Weighted Recall (topic size weighted average over topics) [5] can be used for overall coverage evaluation. Mutual Information [8] and F (cluster size weighted average over clusters) provide overall measures that combine evaluation of quality and coverage.

Although the measurements are reasonable for some kinds of clusterings, they all have problems with overlapping clusters and hierarchical clusterings. Mutual information gives some non ideal clusterings better values than ideal clusterings. When the topics are of very different sizes, Weighted Precision, Weighted Entropy, and F give a high value for useless clusterings (such as a single cluster containing all pages). Average / Weighted Precision and Entropy only measure quality, and are maximized by a set of singleton clusters.

3 New Method - QC4

A fair clustering evaluation method should not inherently favor any particular algorithm. QC4 ensures this by minimizing the bias towards clusterings with particular characteristics (cluster granularity: coarse or fine, clustering structure: hierarchical or flat, disjoint or overlapping): 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.

3.1 The Ideal Clustering

An ideal clustering is created by a human expert based on the pages to be clustered. The classical ideal clustering structure is a single level partition at a chosen granularity. QC4 uses a richer ideal clustering structure to describe clusterings with all kinds of characteristics.

QC4's ideal clustering structure is a hierarchy of topics, organised in levels, so that the set of topics at the top level represents a coarse categorisation of the

pages, and the sets of topics at lower levels represent progressively finer categorisations. This allows QC4 to fairly compare algorithms that produce clusterings of different granularity and to compare algorithms that generate hierarchical clusterings.

Topics may overlap other topics (at the same and different levels), since real pages may belong to multiple topics. However, all pages must be contained in at least one topic at each level. This allows QC4 to evaluate algorithms that return overlapping clusters as well as algorithms that return partitions.

Since search engines often return outliers — pages that are unrelated to all the other pages — the hierarchy may contain a single outlier topic (present at every level) that contains all the outliers. The outlier topic must be disjoint from the other topics. QC4 handles outliers by not counting them when measuring coverage, and by removing clusters that contain a majority of outliers.

3.2 Quality and Coverage Measurements

The standard measures do not work well on hierarchical clusterings with overlapping clusters. Therefore, QC4 introduces four new measures of quality and coverage.

In addition to the notation in section 2.2, 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 level. T_l is the set of topics at level l, T_d is the set of 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.

Cluster Quality Cluster Quality, $QU(c)$, is a measure of how closely a cluster matches a single topic. It is based on a modified entropy measure, $E(c)$.

The standard entropy measure of a cluster does not work with overlapping topics since pages in multiple topics are overcounted. There are two kinds of overlap: overlap of topics at different levels, and overlap of topics at the same level. Overlap between levels is handled by computing the entropy over the topics in a single level. QC4 chooses the level[1], $L(c)$, containing the topic that is the most similar to the cluster as measured by the f-measure.

$L(c) = \text{cluster-level} = lvl(\text{argmax}_{t \in T \setminus T_\emptyset}\{F(c,t)\})$

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

Overlap of topics at the same level is handled by computing a modified precision measure $P'(c,t,t_b)$. The modified measure removes the overcounting by temporarily removing pages in the "best" topic from the other topics, and then normalizing the precision to remove the effect of any other over counting.

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

$E(c)$ measures how focused a cluster is on a single topic, choosing the appropriate level of granularity, and allowing both disjoint and overlapping topics

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

to be handled fairly. However, it does not take cluster and topic size sufficiently into account and it does not recognize random clusters. To account for these, $E(c)$ is scaled down by a new measure that takes account of the cluster and topic size by $S_{recall}(c)$ and recognizes random clusters using $S_{random}(c)$.

$QU(c) = (1 - E(c)) \min\{1, S_{recall}(c), S_{random}(c)\}$

$E(c)$, being a precision/entropy based measure, gives a good value to focused clusters (all their pages belong to the same topic) regardless of the size of the clusters. However, very small clusters, even if they are highly focused, are not very useful to a user if they only contain a small fraction of the topic. To be useful, a cluster should be close to a topic by being both focused on the topic and by being of similar size to the topic. That is, the cluster should not only have good precision/entropy, but should also have good recall. QC4 scales down the quality measure of clusters that are much smaller than the topic that they are focused on by the recall measure. Since a page in a cluster may belong to multiple topics, the standard recall measure was modified to handle pages in multiple topics by averaging the recall of a cluster over all topics weighted by the modified precision $P'(c, t, t_b)$.

$S_{recall}(c) = \max_{t_b \in T_{L(c)}} \{\sum_{t \in T_{L(c)}} P'(c, t, t_b) R'(c, t)\}$

In the web page clustering domain, a cluster with low recall on a small topic is almost useless to the user. On the other hand, a cluster with the same low recall fraction of a very large topic will have more than enough pages for the user to understand the cluster and make an appropriate decision. Therefore, the recall measure can be modified by a non-linear function of the size of the topic to amplify the scaling for clusters focused on small topics.

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

Clusters that are similar to a random selection of pages from the result set provide almost no information, and will not be helpful to the user. Such a clustering should receive near zero quality. However, the modified entropy, $E(c)$, of randomly constructed clusters will generally not be the maximally bad value, especially if the topics are of varying sizes. QC4 uses a modified version of MI, $S_{random}(c)$, to scale down the quality measure of clusters that are similar to a random set of pages. $S_{random}(c)$ has to deal with overlapping topics in a single level, which it does by extracting the intersections of topics into temporary distinct topics and applying MI to the expanded, disjoint set of topics, $\rho(l)$. It also applies a threshold to ensure that only clusters that are very close to random or very small are scaled down. The resulting value is also normalized by the maximum MI to account for the varying maximum value of MI.

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

$S_{random}(c) = \frac{\sum_{r \in \rho(L(c))} |D_c \cap r| log_{|\rho(L(c))|} \frac{|D_c \cap r||D|}{|D_c||r|}}{0.05 \min_{t \in T_{L(c)} \setminus T_\emptyset} \{\sum_{r \in \rho(L(c))} |D_t \cap r| log_{|\rho(L(c))|} \frac{|D_t \cap r||D|}{|D_t||r|}\}}$

Topic Coverage Topic Coverage, $CV(t)$, is a measure of how well the pages in a topic are covered by the clusters. It is an average of the page coverage, $PC(d, t, l)$, of each of the pages in the topic. The coverage uses just level one topics because the page coverage already incorporates topics lower in the hierarchy.

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

A page in a topic is covered to some extent if any cluster contains the page. However, the user is unlikely to find a page if it is in a cluster that appears to be associated with a different topic, so a page will be better covered if it is contained in a cluster that matches a topic that the page is in. The better the match, the better the coverage. If a page is in topic t and cluster c, the precision $P(c,t)$ would be a good measure of how well the page is covered, as long as the page is not also in any other topics or clusters and the cluster is not merely a random selection of the pages. Both topics and clusters can overlap: a page may be in several topics and several clusters. In particular, each page in a top level topic will also be in subtopics of that topic at each level of the hierarchy. Therefore we need something more complicated than precision to measure page coverage.

QC4's page coverage measure considers all the clusters that a page is in, and also all the topics and subtopics the page is in. At each level of the topic hierarchy, it finds the average precision of the clusters that contain the page with respect to the best matching subtopics containing the page. It then recursively computes the maximum of this measure at each level to compute a page coverage measure over the whole hierarchy.

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

$PC'(d,t,l) = \max_{c \in \{c_i | c_i \in C_d \wedge L(c_i) = l\}} \{P(c,t) \min\{1, S_{random}(c)\}\}$

Overall Measurements QC4 has four overall measurements, based on the measures of cluster quality $QU(c)$ and topic coverage $CV(t)$. The overall measurements of clustering quality, AQ and WQ are the average of the cluster qualities, but in WQ they are weighted by cluster size. Similarly, the overall measurements of clustering coverage, AC and WC are the average of the topic coverages, but in WC they are weighted by topic size. The averages give a fairer measure of the smaller, fine grained clusters and topics; the weighted averages give a fairer measure of the larger, broad clusters and topics.

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

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

To compute the overall coverage measures, AC and WC, the topic coverage is averaged over the top level topics of the ideal clustering.

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

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

The measurements fairly evaluate both disjoint and overlapping topics, and topics of varying granularity without bias. Hierarchical and flat clusterings are considered fairly, because the measurements consider the individual clusters, not the hierarchical structure, and cope with overlapping clusters, including clusters that are subsets of other clusters.

4 Evaluation

This section describes how we evaluated QC4 by comparison with existing evaluation measurements. Evaluation of QC4 was completed in two ways: using synthetic test cases and using real world web clustering problems. The synthetic test cases highlight the problem scenarios and boundary cases where existing measurements fail. The real world web clustering tasks show that for simple clusterings, where existing measurements work reasonably well, QC4 reaches conclusions similar to those of existing measurements.

4.1 Synthetic Test Cases

To compare QC4 with existing measurements we devised an extensive set of synthetic test cases. These were organised into eight groups shown in table 1, according to the type of case being tested. The columns of table 1 give the different combinations of evaluation measurements that we considered as overall measurements to compare against QC4, where MI, F, AP, WP, AR, WR, AE, WE are mutual information, f-measure, average precision, weighted precision, average recall, weighted recall, average entropy, and weighted entropy respectively. The tests passed by each overall measurement are shown with a Y in the appropriate rows, for example, QC4 passes all eight tests and the 9th column shows that using just Weighted Precision (Purity) for overall evaluation fails seven of the eight tests.

Table 1. Synthetic test cases comparing QC4 with a wide range of overall evaluation measurements, where Y indicates passing all tests in that rows group of tests

	QC4	MI	F	AP WP AR WR	AE WE AR WR	WP WR	WE WR	AP	WP	AR	WR	AE	WE
Overlapping Clusters	Y	-	-	-	-	-	-	-	-	-	-	-	-
Hierarchical Clusterings	Y	-	-	-	-	-	-	-	-	-	-	-	-
Perfect Clustering	Y	-	Y	Y	Y	Y	Y	Y	Y	-	-	Y	Y
Separate Measures	Y	-	-	Y	Y	Y	Y	-	-	-	-	-	-
Large cluster/topic bias	Y	Y	-	Y	Y	-	-	-	-	Y	-	-	-
Small cluster/topic bias	Y	-	-	-	-	-	-	-	-	-	-	-	-
Random Clustering	Y	Y	-	-	Y	-	Y	-	-	-	-	Y	Y
Split Cluster	Y	Y	-	-	-	-	-	-	-	-	-	-	-

QC4 handles the overlapping and hierarchical clusterings, but none of the other evaluation methods do. QC4 gives perfect scores only to ideal clusterings, but three of the other measures fail; for example, mutual information gives a better than perfect score to a clustering that contains an ideal clustering and a low quality cluster. QC4 includes separate measures for quality and coverage, but

MI and F do not and the individual measures of precision, recall, and entropy do not measure both quality and coverage. QDC handles clusterings with clusters or topics of vastly different sizes where one or more may be relatively large, but eight of the other measures fail; for example, when there is one big cluster containing all pages, the precision, entropy, and weighted recall measures give unduly good scores. QDC handles clusterings with many small clusters or topics, but none of the other evaluation methods do; for example, all other measures give unduly good performance to a singleton clustering (one that has one cluster for each document) and in fact precision, recall, and entropy measures give perfect scores to the singleton clustering. QC4 gives low scores to random clusterings, but seven of the other measures fail; for example, the precision and recall measures can give unduly high scores to random clusterings, often exceeding the scores given to more sensible clusterings. QC4 gives lower scores when perfect clusters are split into smaller clusters, but eleven of the other measures fail; for example, splitting a perfect cluster has no effect of precision, recall, or entropy measures.

The results show that none of the current measurements for overall evaluation are satisfactory, while QC4 passes all tests. While existing measurements can still produce meaningful results and conclusions with simple clustering problems, these tests show that there are conditions under which existing methods can produce inaccurate results, especially with overlapping clusters or hierarchical clusterings. Conclusions drawn from the existing measurements are therefore questionable.

4.2 Real World Web Clustering Tasks

To evaluate QC4 on real world web clustering tasks we selected four queries (Jaguar, Salsa, GP, and Victoria University) and evaluated the performance of nine clustering algorithms (random clustering, and full text and snippet varieties of K-means [1], STC [4], ESTC [5], and QDC [6]) on each of the queries using twelve evaluation measurements (Mutual Information, F-measure and Average and Weighted versions of QC4 Quality, QC4 Coverage, Precision, Recall, Entropy). We averaged the values across the four queries and combined the average and weighted versions of each measurement by averaging them. For the overall evaluation in figure 1C, we also averaged the quality and coverage measures for QC4.

These clustering tasks represented simple clustering problems with little overlap or hierarchy, where existing measurements work reasonably well. Figures 1A, 1B, and 1C show that the QC4 quality, coverages, and overall measures, respectively reach similar conclusions to those of the existing measurements.

In the few cases QC4 differs from the existing measurements, QC4 agrees with the conclusions of the relevant research literature [4–6], which rank the algorithms as QDC, ESTC, STC, K-means, and finally Random clustering, in order of overall web clustering performance. QC4 correctly identifies K-means as a low performing algorithm, whereas F-measure ranks its performance too highly. QC4 correctly identifies ESTC as outperforming STC, whereas mutual

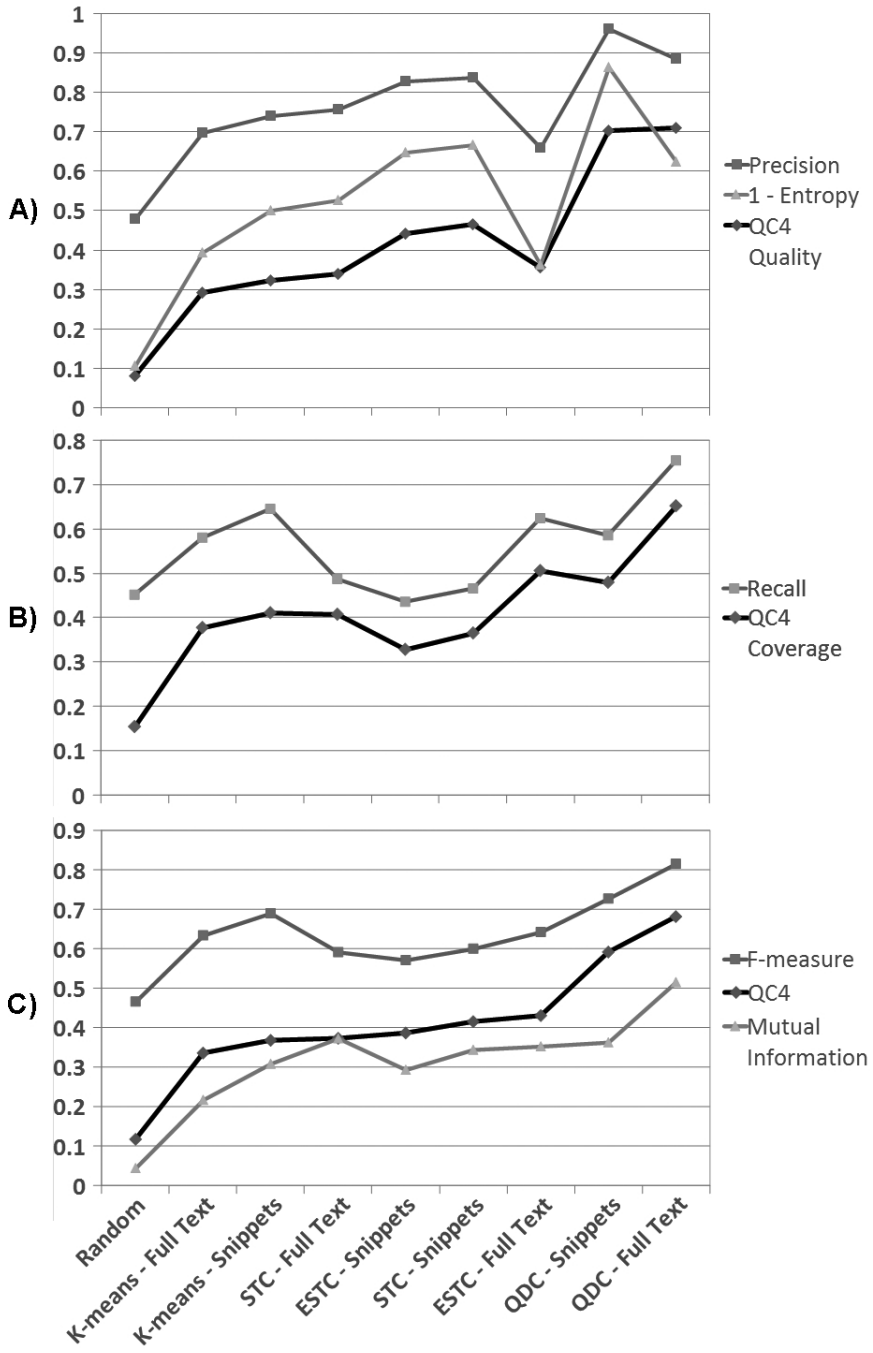

Fig. 1. Comparing measures averaged over four real world web clustering tasks. A) cluster quality measures. B) topic coverage measures. C) overall measures.

information incorrectly identifies STC as the higher performer. This indicates that QC4 makes sensible conclusions on real world tasks.

The real world web clustering tasks also show that QC4 is as expressive as any of the existing standard evaluation methods, and is significantly better than Precision, Recall, and F-measure due to the much lower performance given to random clusterings.

4.3 Applicability to other clustering domains

QC4 has been designed and evaluated with respect to web page clustering, but it can be easily generalized to other clustering domains where clusterings feature different characteristics. The only web specific assumption in QC4 is that it is more desirable to identify small clusters than to extend the coverage of large clusters. If this assumption is not applicable in the clustering domain, the assumption can be removed by simply using the standard recall measure $R(c, t)$ instead of $R'(c, t)$ in QC4's quality measure.

5 Conclusions

This paper introduced QC4, a new clustering evaluation method that allows the fair comparison of all clustering algorithms, even those that produce clusterings with vastly different characteristics (cluster granularity: coarse or fine, clustering structure: hierarchical or flat, disjoint or overlapping, and cluster size: large or small). QC4 achieved this by generalizing the gold-standard approach to use a more general ideal clustering that can describe ideal clusterings of varying characteristics and introduced four new overall measurements that function with clusterings of different characteristics fairly in terms of cluster quality and topic coverage.

QC4 was evaluated by comparison to the standard evaluation measurements in two ways: on an extensive set of synthetic test cases and on a range of real world web clustering tasks. The synthetic test cases show that QC4 meets all the requirements of a good evaluation measurement, while all the current measurements fail with overlapping clusters, hierarchical clusterings, and some boundary cases. On simple real world web clustering tasks, where the existing methods are less affected by the conditions tested by the synthetic test cases, the results show that QC4 is at least as good as the existing evaluation measurements and gives a better evaluation in several cases.

In the future, standard test data sets can be constructed and used to evaluate standard clustering algorithms to provide a baseline for comparison. QC4 should also be evaluated on other clustering domains, especially those where clusterings have different characteristics.

Acknowledgements

Daniel Crabtree is supported by a Top Achiever Doctoral Scholarship from the Tertiary Education Commission of New Zealand and was supported during this research by an SMSCS Research Grant.

References

1. Steinbach, M., Karypis, G., Kumar, V.: A comparison of document clustering techniques. In: KDD Workshop on Text Mining. (2000)
2. Ali, R., Ghani, U., Saeed, A.: Data clustering and its applications. `http://members.tripod.com/asim_saeed/paper.htm` (1998)
3. Wang, Y., Kitsuregawa, M.: 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. (2002) 499–506
4. Zamir, O.E.: Clustering Web Documents: A Phrase-Based Method for Grouping Search Engine Results. PhD thesis, University of Washington (1999)
5. Crabtree, D., Gao, X., Andreae, P.: Improving web clustering by cluster selection. In: The 2005 IEEE/WIC/ACM International Conference on Web Intelligence. (2005) 172–178
6. Crabtree, D., Andreae, P., Gao, X.: Query directed web page clustering. In: The 2006 IEEE/WIC/ACM International Conference on Web Intelligence. (2006)
7. Halkidi, M., Batistakis, Y., Vazirgiannis, M.: On clustering validation techniques. Journal of Intelligent Information Systems **17**(2-3) (2001) 107–145
8. Strehl, A.: 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)
9. Tonella, P., Ricca, F., Pianta, E., Girardi, C., ITC-irst, Lucca, G.D., Fasolino, A.R., Tramontana, P., di Napoli Federico II, U., Napoli, Italy: Evaluation methods for web application clustering. In: 5th International Workshop on Web Site Evolution, Amsterdam, The Netherlands. (2003)
10. Meila, M.: Comparing clusterings. Technical Report 418, Department of Statistics, University of Washington (2002)
11. chiu Wong, W., Fu, A.: 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. (2000)
12. Crabtree, D., Gao, X., Andreae, P.: Standardized evaluation method for web clustering results. In: The 2005 IEEE/WIC/ACM International Conference on Web Intelligence. (2005) 280–283
13. van Rijsbergen, C.J.: Information Retrieval. Butterworths, London (1979)
14. Mackay, D.J.: Information Theory, Inference, and Learning Algorithms. Cambridge University Press (2003)