Standardized Evaluation Method for Web Clustering Results

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

Abstract

Web clustering assists users of a search engine by presenting search results as clusters of related pages. Many clustering algorithms with different characteristics have been developed: but the lack of a standardized web clustering evaluation method that can evaluate clusterings with different characteristics has prevented effective comparison of algorithms. The paper solves this by introducing a new structure for defining general ideal clusterings and new measurements for evaluating clusterings with different characteristics by comparing them against the general ideal clustering.

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 is often huge, and many of the results may be irrelevant to the user's goal.

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. Many clustering algorithms have been developed including: K-means [6], Hierarchical Agglomerative Clustering [1], Link and Contents Clustering [10], and Suffix Tree Clustering [11]. Many different evaluation methods and measurements [7, 3, 8, 11, 6, 5, 2] have been used to evaluate web clustering algorithms and the results are often incomparable.

A standardized evaluation method for comparing web clusterings is difficult because different algorithms produce clusterings with different characteristics: the clustering granularity may be coarse or fine; the clusters may be disjoint 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 of the existing evaluation methods are biased towards algorithms that produce clusterings with certain characteristics.

This paper presents a new standardized evaluation method that enables the evaluation and comparison of web clusterings with different characteristics by generalizing the "gold-standard" approach to use a new structure for ideal clusterings and by developing new measures of quality and coverage. The new evaluation method is standardized in the sense that it allows the fair comparison of all web clustering algorithms, even those that produce clusterings with vastly different characteristics.

The next section outlines previous methods and measurements. Section 3 describes our proposed method. Section 4 discusses the evaluation of our new method. Section 5 concludes the research and provides direction for future work.

2. Previous Methods and Requirements

2.1. Previous Methods

There are two broad methodologies for evaluating clusterings. Internal quality [7, 3] evaluates a clustering only in terms of a function of the clusters themselves, with no reference to external information about the desired output. External quality [7, 3] evaluates a clustering using external information, such as an ideal clustering. Where there is such external information, 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 [8], task-oriented [8], and user evaluation [11]. Gold-standard approaches manually construct an ideal clustering with each ideal cluster la-

0-7695-2415-X/05 $20.00 © 2005 **IEEE**

beled with a topic, 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 of the clustering for users.

Task-oriented methods such as search result reordering have a bias towards the selected task, making them poor candidates for a standardized clustering evaluation method. User evaluation methods are very difficult to reproduce and the large cost, and time involved in conducting good 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. We refer to the clusters of the ideal clustering as topics (T), to distinguish them from the clusters (C) of the clustering being evaluated.

A clustering is perfect if its clusters match the topics. 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 evaluation methods must measure both factors.

The Purity [7] and F measurements [6, 7, 2] are based on precision and recall [9]. 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 [9], F(c,t), combines P(c,t) and R(c,t). The Purity of a clustering is the average precision of the clusters relative to their best matching topics, and F is the average F-measure of the clusters relative to their best matching topics.

The Entropy and Mutual Information measures [6, 7] are based on information theory [4]. 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 pair of a cluster and a topic.

Overall measurements used by the current methods are not satisfactory. Firstly, they do not measure coverage well. Entropy and Purity only measure quality, and are also biased towards small clusters (and maximized by a set of singleton clusters). Also, when the topics are of very different sizes, Entropy, F, and Purity give a high value for useless clusterings such as a single cluster containing all pages, and are biased by the performance of large clusters. MI, and F include coverage, but combine it with quality, making it impossible to interpret the two factors separately. MI also requires that both the clusters and the topics partition the pages.

3. New Method - QC4

QC4 (Quality, Coverage, and 4 Overall Measurements) addresses the problem of an overly constrained ideal clustering by introducing a new structure for representing more flexible ideal clusterings and new overall measurements that fairly characterize clusterings in terms of quality and coverage.

3.1. A New 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. Our proposed 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

QC4[1] generates four overall measurements, based on measures of cluster quality $QU(c)$ and topic coverage $CV(t)$. The overall measurements of the quality

[1] Detailed formulas for all measures can be found in our technical report at http://www.mcs.vuw.ac.nz/comp/Publications/

of a clustering are the average of the cluster qualities (AQ), and the average of the cluster qualities weighted by cluster size (WQ). Similarly, the overall measurements of the coverage of a clustering are the average of the topic coverages (AC), and the average of the topic coverages weighted by topic size (WC). The averages give a fairer measure of the larger broad clusters and topics; the weighted averages give a fairer measure of the smaller fine grained clusters and topics. In evaluating a web clustering algorithm for a particular application, a single appropriately weighted combination of the four overall measurements should be used.

3.2.1. 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 measurement, $Q(c)$, scaled down by a factor identifying problematic clusters and topics.

The standard entropy measure of a cluster c is the average (over the topics) of log precision: $-\log(P(c,t))$. However, entropy 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 average log precision over the topics in a single level. QC4 chooses the level containing the topic, t_{max} that is the most similar to the cluster as measured by the f-measure $F(c,t)$.

Overlap of topics at the same level is handled by computing a modified precision measure $P'(c,t)$. 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. The best topic is the one that maximises $Q(c)$. $Q(c)$ is computed from

$$-\sum_t P'(c,t)\log(P'(c,t))$$

where the sum is taken over the topics in the level.

$Q(c)$ measures how focused a cluster is on a single topic, choosing the appropriate level of granularity, and allowing both disjoint and overlapping topics to be handled fairly. However, it does not take cluster and topic size sufficiently into account and it does not recognize random clusters.

$Q(c)$, being a precision/entropy based measure, gives a high measure 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 clusters that are much smaller than the topic that they are focused on. 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 $P'(c,t)$.

A cluster with low recall on a small topic will have very few pages, and therefore will be 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 modified recall measure is further modified by a non-linear function of the size of the topic to amplify the scaling for clusters focused small topics.

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, modified entropy $Q(c)$ of randomly constructed clusters will generally not be zero, especially if the topics are of varying sizes. QC4 uses a modified version of MI to identify clusters that are similar to a random set of pages, and then scales down $Q(c)$ appropriately. The modified version of MI has to deal with overlapping topics in a single level, which it does by extracting the intersections of topics into distinct topics, and applying MI to the expanded disjoint set of topics. It also applies a threshold to ensure that only clusters that are very close to random or that are very small are scaled down.

3.2.2. Topic Coverage Topic Coverage, $CV(t)$, is a measure of how well the pages in a topic are covered by the clusters. It is computed from an average of how well each of the pages in the topic are covered.

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. Because topics and clusterings can overlap, a page may be in several topics and several clusters, and therefore we need something more complicated than precision to measure page coverage. In particular, each page in a top level topic will also be in subtopics of that topic at each level of the hierarchy.

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

IEEE COMPUTER SOCIETY

at each level to compute a page coverage measure over the whole hierarchy.

To compute the overall coverage measures, AC and WC, the topic coverage is averaged over just the top level topics of the ideal clustering. There is no need to compute the topic coverage of lower level topics since each top level topic implicitly includes topics below it in the hierarchy.

4. Evaluating QC4

We analysed the output of QC4 on a variety of synthetic data sets and ideal clusterings that demonstrate a number of extreme properties that are not handled by other evaluation methods. The synthetic test cases analysed included the following: perfect clustering, near perfect clustering, set of all singleton clusters, mid-sized cluster containing a near uniform distribution of pages from topics, overlapping clusters, and several distinct combinations of clusters at different levels of granularity. We explored the effect of a variety of possible clusterings and compared them using QC4. In all cases, QC4 generated reasonable measures that preferred the clusterings that appeared to be more useful to a user. On the other hand, the other evaluation methods produced inappropriate or nonsensical evaluations.

Clearly, we need to apply QC4 to real data sets to determine whether it generates useful and believable measures on a variety of clustering algorithms, including algorithms that generate overlapping clusters, hierarchies of clusters, and clusters at different levels of granularity. We are currently performing such experiments on real data sets, and the preliminary results are encouraging.

5. Conclusions

This paper introduced QC4, a new standardized web clustering evaluation method. QC4 minimizes method bias by generalizing the commonly used gold-standard approach to use a more general ideal clustering that can describe multiple 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. Our analysis has shown that QC4 significantly outperforms MI on many synthetic test cases that cover a broad range of clustering characteristics.

We are currently investigating the performance of QC4 on several real data sets and a variety of clustering algorithms.

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. The software clustering methods for aiding the ideal clustering construction process could be adapted to aid the construction of ideal web clusterings for QC4.

Acknowledgements

Daniel Crabtree is supported by a Top Achiever Doctoral Scholarship from the Tertiary Education Commission of New Zealand.

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

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

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

[7] 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.

[8] 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.

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

[10] 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.

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