Single-Pass Clustering for Peer-to-Peer Information Retrieval: The Effect of Document Ordering

Iraklis A. Klampanos
Department of Computing Science,
University of Glasgow, Scotland.
Email: iraklis@dcs.gla.ac.uk

Joemon M. Jose
Department of Computing Science,
University of Glasgow, Scotland.
Email: jj@dcs.gla.ac.uk

C. J. "Keith" van Rijsbergen
Department of Computing Science,
University of Glasgow, Scotland.
Email: keith@dcs.gla.ac.uk

***Abstract*— Document clustering has been a particularly active research field within the Information Retrieval (IR) community. Among the numerous clustering algorithms proposed, single-pass clustering stands out in terms of both time and space efficiency. However, it is generally acknowledged that single-pass clustering has a major defect, namely its output depends on the order in which documents are presented. Building on our previous work, and having identified single-pass clustering as potentially useful for P2P IR, we study the extent to which this is true in practical terms. We do so by experimenting with two large web-based testbeds, which are suitable for Peer-to-Peer IR evaluation. The results of our study show that document ordering does not practically matter for single-pass clustering.**

I. Introduction and Motivation

Document clustering has been a particularly active research field within the Information Retrieval (IR) community [1], [2], [3], [4]. Organising documents according to their content, and consequently, achieving more accurate and effective retrieval is, arguably, one of the principal goals of IR research. The reason behind this, apart from a natural human tendency [5], is that by clustering, documents relevant to the same topics tend to be grouped together (the Cluster Hypothesis – [1]).

Single-pass clustering is one of the incremental clustering algorithms, and requires only one pass over the document descriptions to be clustered [1]. Single-pass clustering algorithms do not possess the desirable theoretical properties of other clustering techniques, such as hierarchical algorithms. In particular they are known to be affected by the order in which the documents are presented to the algorithm. However, their space and time complexity (please refer to Section II for more details) as well as their incremental operation are particularly attractive for a range of different applications involving clustering streams of documents.

Clustering has been used in many P2P IR solutions such as [6], [7], [8], [9] and variants of single-pass clustering are expected to be employed in future systems due to their simplicity and efficiency. We have identified and effectively used the single-pass clustering scheme as part of a P2P IR architecture [6]. However, in P2P IR systems, most peers have initial sets of documents which are used to bootstrap the network. It is important to see in which way document ordering affects the effectiveness of single-pass clustering. Since the clusters that get generated during the initial steps of the algorithm are then used as bins in which consecutive documents are placed, it would be beneficial to know how to reach the best possible initial classification.

We make the assumption that single-pass clustering makes more informed decisions the more documents it has processed (please refer to Section II for the exact algorithm studied), even more so if the initial clusters that have been identified are more accurately formed. Therefore, if there is little cost involved in re-ordering the bootstrapping sets of documents of the participating peers and if there is a document ordering which results in more effective document classifications, overall retrieval effectiveness can be potentially increased. This is the main motivation behind this study, which looks into different orderings that could aid a simple single-pass algorithm to create more retrieval-effective clusters.

An additional aspect we are interested in investigating is the usefulness of single-pass clustering in terms of retrieval effectiveness, within the context of P2P IR. The motivation behind this is the general belief that single-pass algorithms do not perform well. We believe it is informative to study the extent at which single-pass clustering algorithms do or do not stand up to expectation, and argue about their usefulness with respect to possible retrieval effectiveness rather than to their algorithmic, time and space, properties.

The layout of this paper is as follows: In the next section, we briefly review a number of P2P IR architectures that have used clustering procedures. We will also describe a P2P IR architecture which uses single-pass clustering. In Section III we introduce the single-pass algorithm as well as the different document orderings we have experimented with. In Section IV we describe the experimental methodology we followed and we also present the findings of our experiments. Following that, in Section V we provide a discussion on our findings and, finally, in Section VI we present our conclusions and discuss future work.

II. Background: Clustering in P2P IR

Clustering, based on networking properties or content, has been employed or assumed to exist in a number of studies or systems dealing with P2P IR and related areas.

In [8], Krishnamurthy et. al. employ network-aware clustering, dealing primarily with file-sharing problems by describing each file as a set of filename keywords. In this solution, peers get clustered through a central clustering service. In

INFOSCALE '06. Proceedings of the First International Conference on Scalable Information Systems, May 29-June 1 2006, Hong Kong

[9], Khambatti et. al. refer to communities of peers seen as interest groups based on sets of attributes. These attributes are either set manually by the users of the system, or derived automatically from past queries. Additionally, in [10], Ng and Sia base their multimedia retrieval around a phone-directory-like networking structure.

A. A Hybrid P2P IR Network

Concerning content-based P2P IR, in [7], Lu and Callan assume a clustering of leaf nodes of similar topic areas around *directory nodes* (super-peers that are responsible for the routing of queries and results), working within the context of digital libraries. In their study, directory nodes were expected to be covering specific types of content. In order to translate this into an evaluation testbed, they had to cluster a large part of TREC's WT10G collection and then to assign each cluster (content region) to each directory node. Even though clustering was not an integral part of this solution, it proved crucial for its evaluation; it was an assumption taken on the information environment of their system.

B. A Cluster-Based P2P IR Architecture

In a previous study [6], we identified clustering as a potentially beneficial technique for the automatic organisation of a content-based P2P IR network. This architecture has been proposed by considering many practical applications. Clustering takes centre stage in this solution. It forms the basis of resource description and, therefore, also affects resource selection.

In this architecture, each peer that shared documents first has to cluster its local collection. Then, it has to inform a nearby *hub* (super-peers similar to Lu and Callan's directory nodes in terms of responsibilities) of its content by passing it the descriptions of its collection's clusters. Hubs, then, organise the peers into content-aware peer groups, based on these clustering descriptions. Each of these peer groups gets described by a single vector, usually the average of its constituent vectors.

Query dissemination takes place in two steps. First, the query gets compared to the peer-group descriptions and gets sent to the closest ones. During the second step, once a query has reached a peer group it is compared against the peers' descriptors and gets sent to the closest ones (according to a similarity function). It is obvious that in such a P2P setting the effectiveness of the clustering methods used has a great impact on the overall retrieval effectiveness. However, the dynamic nature of the information-sharing peers also pinpoints the need for an incremental clustering solution, hence the importance of single-pass clustering in P2P IR.

We presented only a fraction of P2P IR research that attempts to organise the network in order to improve retrieval effectiveness and performance. Therefore, we believe it is beneficial to study how an efficient online clustering algorithm works, and that the findings of our study might find application in new cluster-based systems or the evaluation of existing ones.

III. The Single-Pass Clustering Algorithm

Single-pass clustering, as the name suggests, requires a single, sequential pass over the set of documents it attempts to cluster. The algorithm classifies the next document in the sequence according to a condition on the similarity function employed. At every stage, the algorithm decides on whether a newly seen document should become a member of an already defined cluster or the centre of a new one. In its most simple form, the similarity function gets defined on the basis of just some similarity (or alternatively, dissimilarity) measure between document-feature vectors.

In our experiments the similarity function is the cosine coefficient applied to term-frequency vectors, while the description of a cluster is the average vectors of the documents included in the cluster in question. The exact algorithm we studied is the following:

1) **for each** document d in the sequence **loop**
   a) find a cluster c that maximises $\cos(c, d)$;
   b) **if** $\cos(c, d) > t$ **then** include d in c;
   c) **else** create a new cluster whose only document is d;
2) **end loop.**

In this algorithm, t is the similarity threshold value, which is usually derived experimentally.

While this algorithm, especially in this simple form, is straightforward to implement and apply, it has some serious theoretical drawbacks [1]. First, the very fact that the output is known to be crucially dependent on the order in which documents are input makes it difficult to model mathematically. Second, as van Rijsbergen points out, "the effects of errors in the object descriptions are unpredictable" [1].

However, the time complexity of this algorithm is very attractive, $O(n \log n)$, and its space complexity is $O(n)$, where n is the number of documents to be clustered. Additionally, there are many cases, P2P IR being one of them, where documents are expected to be continuously arriving into or getting removed from a collection. In these cases, clustering has to take place on-the-fly, without any prior, global knowledge of the collection's properties. The objective of this research is to study the effect of document ordering in single-pass clustering within a P2P context. We study a number of document orderings, which are presented in Section III-B.

A. Experimental Testbeds

For experimenting, we used two testbeds which have been shown to be suitable for evaluating P2P IR systems [11]. They are based on TREC's WT10G collection, a web-based collection containing 1.69 million documents, spanning 11,680 web domains [12].

The first testbed comprises of the web domains present in WT10G, kept unchanged and representing one peer-collection each (ASISWOR). In the other testbed the documents are uniformly distributed across as many sub-collections as there are domains, based on some criteria (UWOR). Please refer to [11] for more information about these WT10G partitions as

well as the reasoning behind their potential usefulness in P2P IR evaluation.

By choosing these testbeds we, first, wanted to see how single-pass clustering behaves with documents that are of lesser quality than proper articles, for instance, documents present in other TREC collections like AP, WSJ etc. Second, we wanted to study the effect of document ordering on single-pass clustering in document environments that could be found in P2P IR settings.

B. Document Orderings

From the algorithm presented in the previous section, it appears that the later runs of the algorithm are better informed than the initial ones. By claiming that they are better informed we mean that there is a larger pool of clusters to choose from when looking for the best candidate for a document d. The final classification obtained can be altered by changing the order in which documents are fed into the algorithm. Since the similarity method used is applied on the term frequency vectors of the documents, in order a new ordering to be effective, it has to be based on the term-related properties of the documents as well.

In this section we present the different orderings we tested and we justify their usefulness in our study.

a) Random (RAND).: This is just a random ordering of the documents in the collection, regardless of their sizes. The classification obtained by using this ordering was thought of as a baseline in this study. This is a different ordering from the "natural" order that documents appear in TREC's WT10G. We chose this intentionally, in order to differentiate between a truly random ordering and some ordering which, although independent of term-frequency or document-length properties, is possibly based on date information and/or the positioning of a document within its domain's directory tree.

b) Inverse-Unique-Terms-Ordering (IUTO).: This is the ordering of the documents according to the number of their unique indexing terms. In this setup, documents with the fewest unique indexing terms are positioned at the head of the queue, while documents with the most unique indexing terms are positioned at the tail of the queue. This arrangement was expected to result in a very poor classification – in fact the worst case among the orderings we present in this section. This is due to the fact that the initial clusters get created based on very little information, that is just based on the few terms contained in the smallest documents. This ordering was also introduced as a baseline and was expected to result in interesting classifications especially when compared to the random document ordering presented above.

c) Length-Ordering (LO).: This is the ordering of the documents according to their length, regardless of whether they have fewer or more unique indexing terms. The largest document comes at the head of the queue while the smallest comes at the tail. This ordering was considered under the assumption that larger documents are more important, and, therefore, might be more appropriate as centres of newly created clusters. Hence, it was expected to lead to better classifications than both the random and the inverse ordering based on unique indexing terms.

d) Unique-Terms-Ordering (UTO).: This is the reverse ordering to the *IUTO* ordering mentioned above. In this ordering, the document with the largest number of unique terms is positioned at the head of the queue. This ordering shares the assumption that justified the use of length-ordering. However, the difference is that length alone does not make a document more suitable to be the centre of a cluster if this is caused by great repetition of a few terms. This ordering was expected to give better classifications than the length-based ordering, since the decision of assigning a document to a cluster should become better informed, early in the algorithm's running time.

e) Complementary-Documents-Ordering (CDO).: This ordering is based on the assumption that the earlier the algorithm will deal with all terms in the collection's vocabulary, the more likely it is for its decision to be accurate, and the better the classification it will produce. So, in this ordering, the fewest documents that span the collection's vocabulary are placed at the head of the queue. The rest of the documents are placed according to their length, with the bigger documents coming first. This ordering was expected to lead to better classifications than all the previously mentioned orderings.

IV. Experimentation

A. Methodology

In these experiments, we clustered each sub-collection by using the single-pass clustering algorithm presented in Section III. For the evaluation of the clusters we used the relevance assessments provided with the WT10G collections, which are based on 100 topics. Following the Cluster Hypothesis [1], we would expect documents that are relevant to the same topic to end up in the same cluster. Looking into each sub-collection individually, we computed precision and recall for each cluster that contained at least one relevant document. In other words, we treated these, "relevant", clusters as resulting sets of documents having to do with particular topics. For the sake of clarity, we will be referring to these values as *cluster precision*, CP, and *cluster recall*, CR. Finally, the harmonic mean, E, was computed between these values giving equal weight to CP and CR. The exact calculations are the following:

$$CP = \frac{\text{\# of relevant documents in cluster}}{\text{\# of documents in cluster}}$$

$$CR = \frac{\text{\# of relevant documents in cluster}}{\text{\# of relevant documents in domain}}$$

$$E = 1 - \frac{1}{0.5/CP + 0.5/CR}$$

The measures of CP, CR and their harmonic mean are a useful tool in assessing the quality of a given cluster. Single-pass clustering is known to generate very diverse clusters and the E measure seems to be handling the extremes, of too

large and too small clusters, very well. For instance, within a specific domain, a cluster of size 1, whose document is relevant to a given topic, is only very good if there are no other clusters with relevant documents to that topic. If this is the only relevant document within the given domain, E would become equal to 0, which is the best obtainable. This means that the relevant documents present in the domain are concentrated in one, single cluster. If, on the other hand, a topic has a number of relevant documents scattered across many clusters within a domain, E becomes larger, indicating a classification of lesser quality.

An additional aspect we explored was the distributions of documents that are relevant to any of the topics. We observed how these distributions change in clusters of various levels of E. This is an important indication of whether single-pass clustering is a feasible solution for P2P IR from the perspective of routing queries to appropriate resources. It essentially answers the question of whether we would reach a satisfactory level of recall by choosing to route a query to the top-ranked set of peers by applying a measure such as the cosine similarity coefficient.

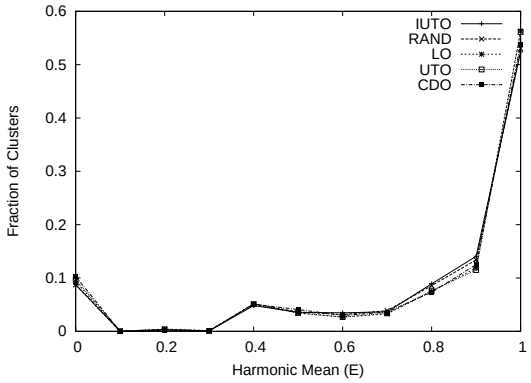
B. Results

Figures 1(a), 1(b) and 1(c) depict the distributions of relevant clusters (i.e. the clusters that contain at least one relevant document to one of the TREC topics) of the ASISWOR sub-collections (the Web-domains). Figures 2(a), 2(b) and 2(c) present the distributions of relevant clusters of the UWOR sub-collections. In both figures it is shown how the choice of a different threshold value (t in the algorithm of Section III) affects these distributions of clusters, but not the clusters' effectiveness depending on document ordering[1].

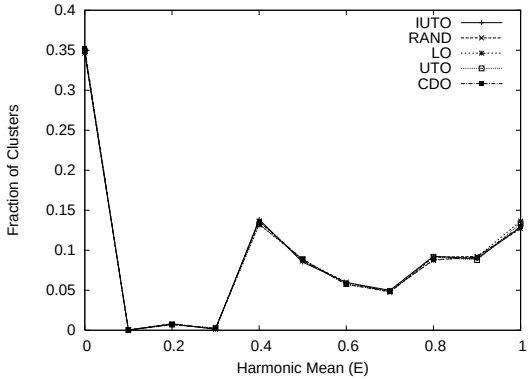
The results that are relevant to the other part of our study, i.e. how relevant documents are covered in clusters of different levels of E, can be seen in Figures 3 and 4 for the ASISWOR and the UWOR testbeds respectively. Again, we present the cases of three different thresholding values (0.3, 0.6 and 0.9) which we used when we clustered the sub-collections. These graphs were compiled by calculating the percentages of the relevant documents contained within the clusters of given levels of E (i.e. the number of relevant documents within a cluster over the total number of relevant documents for each topic). Again it appears that document ordering does not make any difference.

In Figures 1 and 2, the x-axis represents a number of bins starting with this of $E = 0$, containing only clusters that achieve $E = 0$, on the left. Every consecutive bin contains the fraction of clusters achieving an E value that ranges from the E value signified by the bin on its left up to the current bin's signified value (e.g. at point $x = 0.4$ we see the fraction of clusters that achieve an E value from 0.3, exclusive, up to 0.4). On the other hand, in Figures 3 and 4, the x-axis represents again a number of bins, however, in these graphs

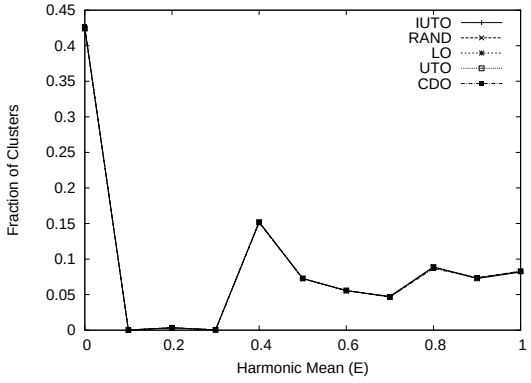
[1]When analysing our results, we did not perform any significance testing, because of the complexity of the procedures we followed.



(a) 0.3

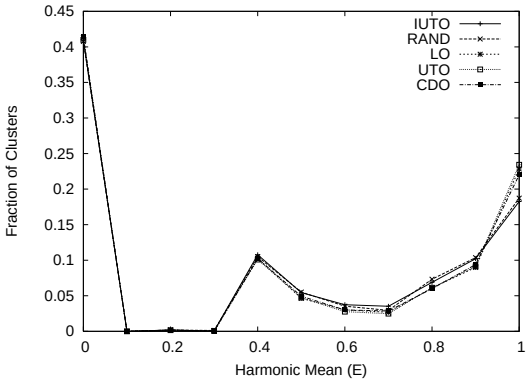


(b) 0.6

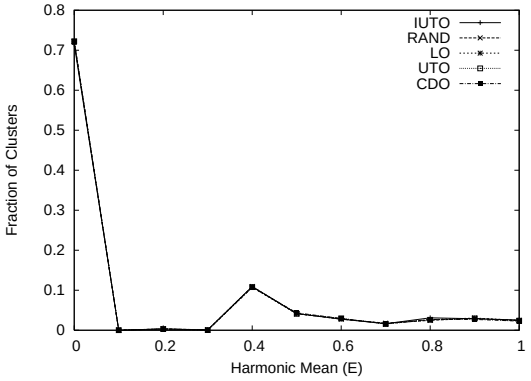


(c) 0.9

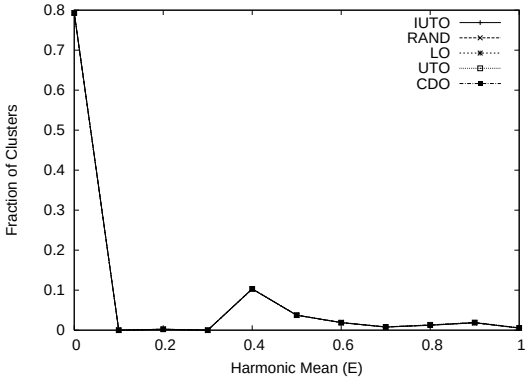
Fig. 1. The distribution of the relevant clusters (created in the Web domain sub-collections) according to their E value.



(a) 0.3

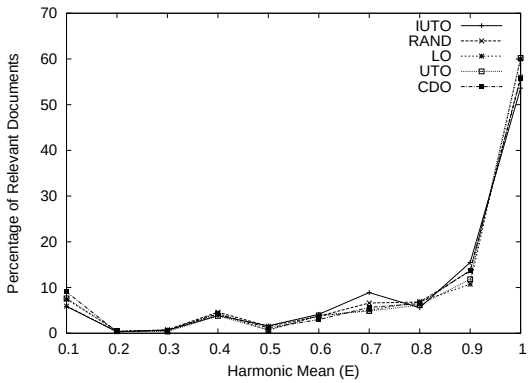


(b) 0.6

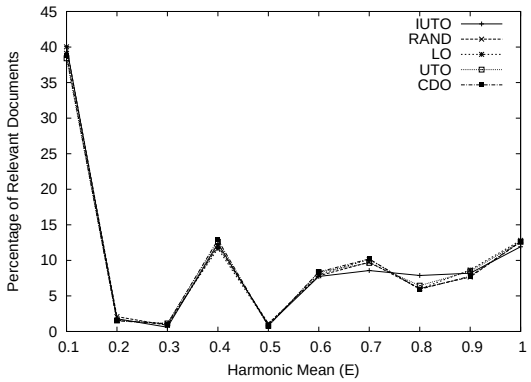


(c) 0.9

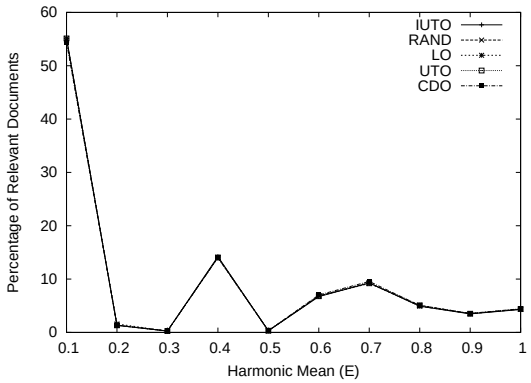
Fig. 2. The distribution of the relevant clusters (created within the uniform collections) according to their E value.



(a) 0.3

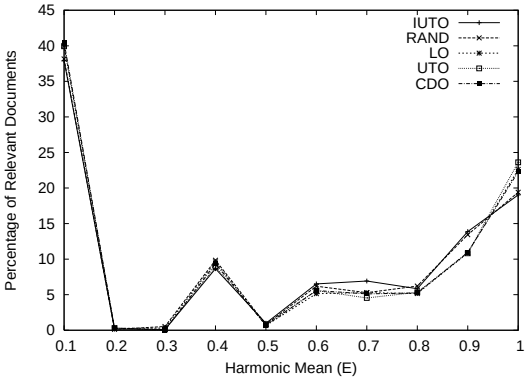


(b) 0.6

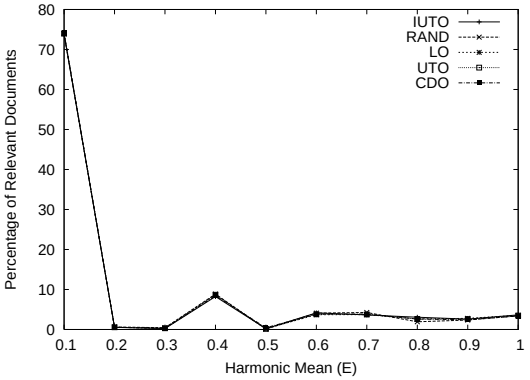


(c) 0.9

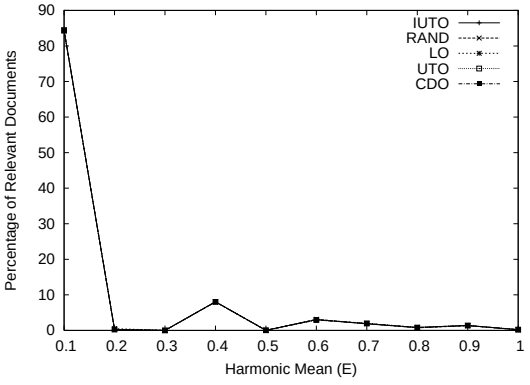
Fig. 3. The percentages of relevant documents in all topics, at different levels of E for the ASISWOR testbed.



(a) 0.3



(b) 0.6



(c) 0.9

Fig. 4. The percentages of relevant documents in all topics, at different levels of E for the UWOR testbed.

every point represents a bin of E values that starts at the value signified by its immediate point on the left and ranges up to its signified value, i.e. these graphs do not treat $E = 0$ as a special case. This discrepancy was left intentionally in order to observe and emphasise the fraction of "perfect" clusters in the first part of the experiments.

The experimental results clearly indicate that document ordering does not actually matter. Of course, this claim holds experimentally, and not theoretically, and for collections that share similar characteristics to those of web domains, i.e. for collections of documents of various lengths and of questionable quality. At this point we also ought to clarify that, even these results show a remarkable consistency at various thresholds (conditioning on the cosine similarity coefficient), they only hold for certain for the distribution of the relevant documents of the standard 100 topics provided by TREC.

On the other hand, and in defence of using these Web-based collections, it should be mentioned that single-pass clustering would be expected to behave in this way in more consistent collections like these consisting of properly written and managed articles. We base this claim on the fact that documents in such collections are more homogeneous with respect to their length and their written qualities (vocabulary, typographical errors etc.). Therefore, irrespectively of the order in which they are presented to an online clustering algorithm, such as the one studied herein, they should lead to better classifications, since almost each document could make a suitable candidate for the algorithm to start with. Furthermore, we managed to replicate our results in two different sets of collections, in which documents are distributed differently. In both cases, even though the distribution of relevant documents is important to the final quality of the classification, it does not seem to be making any difference to the behaviour o different document orderings for single-pass clustering.

V. Discussion

In this section, we discuss the results presented in the previous Section while we also attempt to reason about some of their evident properties. We do so by, first, stating that the results we obtained show emphatically that document ordering *does not* have any practical impact on classifications resulting from single-pass clustering.

A. The Distributions of Relevant Clusters

For the case of Web domains (ASISWOR – Figure 1) and for a threshold value of $t = 0.3$, about 10% of the clusters appear to be containing solely relevant documents to the topics. The corresponding figure for the collections containing uniformly distributed documents (UWOR – Figure 2) is slightly higher than 40%. These graphs also show a peak for at the percentages of clusters that are of the worst-quality ($0.9 < E \leq 1$). These figures are about 55% for the clusters of ASISWOR and around 20% for the clusters of UWOR. The rest of the clusters appear to be distributed, almost uniformly, at rates a little lower than 10%. An exception can be observed at bins of $0.1 < E \leq 0.3$ from which relevant

clusters are virtually absent. This last observation is rather disappointing considering that these clusters would also be potentially important for discovery in a distributed setting such as a P2P network.

As the threshold increases we observe a transfer of the spike from the far-worst E bins to the best. Again, with the exception of the bins representing $0.1 < E \leq 0.3$, the rest of the clusters appear to be, more or less, uniformly distributed at percentages around and lower than 10%, for both cases of collections (ASISWOR and UWOR). We believe this could have happened due to the larger number of clusters we obtained for higher threshold values, and therefore due to the greater partitioning that occurred in biggest groups of documents.

An additional artifact that these graphs show, as the clustering threshold grows, is the seemingly less difference that ordering seems to be making to the final classifications. While for the case of $t = 0.3$ we can observe some difference in the histograms, for the case of $t = 0.9$ the graphs seem to touch perfectly. This is again due to the fact that, for higher threshold values, the partitioning of bigger groups of documents becomes more violent, resulting in numerous tiny clusters, and therefore, making their numbers for the same E values to be more likely to coincide.

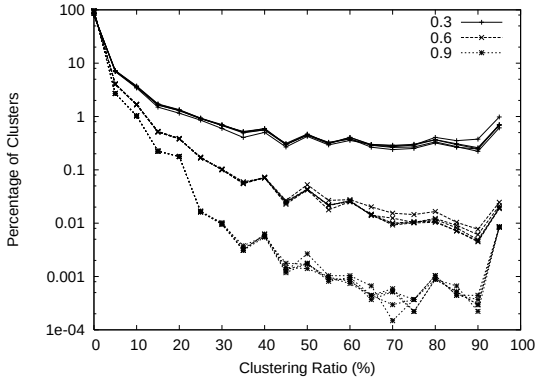
B. The Distributions of Relevant Documents

This reasoning, having to do with large numbers of small clusters, gets reinforced by the second part of our experiments. In Figures 3 and 4 we can see the distribution of relevant documents within clusters of various E values.

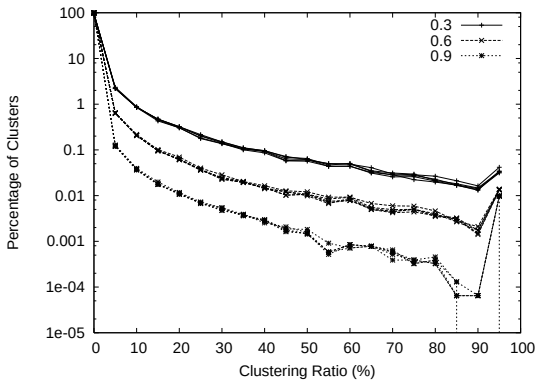
The first observation we can make is that document ordering does not appear to matter as far as the percentages of relevant documents are concerned either; at least not for the best-quality clusters (of E values closer to 0), which would be more preferable for query routing at a distributed IR setting. Furthermore, these figures resemble the previous graphs at a great extent, which can only mean that the number of very small clusters, even singletons (clusters containing a single document) is very high. Indeed, it must mean that the number of extremely small clusters is so high that any theoretically potential significance that document ordering might imply gets diminished. However, this does not only appear to be happening for the higher thresholding values of 0.6 and 0.9, but also for the more modest $t = 0.3$.

An additional contributing factor to this might be the quality and properties of the documents we examined. It might be the case that the similarity between web documents is difficult to exceed very low values because of the restricted vocabularies, possible typographical errors and other adverse artifacts.

In order to make sure that this assertion holds, we looked into the clustering ratios of the clusters we created for both testbeds. By the term clustering ratio, we are referring to the number of documents residing in each cluster divided by the total number of documents of its collection. This can also be thought as a kind of compression ratio, in the sense of what percentage of a single collection can be represented by each



(a) ASISWOR



(b) UWOR

Fig. 5. The distributions of clusters according to their clustering ratios, for the two sets of collections.

of its clusters. The distributions of clusters according to their clustering ratios, for the two sets of collections and for all document orderings and threshold values can be seen in Figure 5. Please note that the y-axis has been plotted in logarithmic scale.

In this figure it can be seen that, even for a threshold value of 0.3, almost 100% of the clusters obtained by any document ordering and in the collections of both testbeds have clustering ratios that range between 0% and 5%. This, meaning that the overwhelming majority of clusters are of very small size, explains the lack of difference between any two document orderings in terms of effectiveness. It also implies that single-pass clustering failed to discover clusters of documents in the collections used from the two testbeds, always assuming that a classification of documents was there to be discovered in the first place.

VI. Conclusion and Future Work

In this paper we studied the effect that different document orderings have on single-pass clustering. We explored five different document orderings while clustering two testbeds

comprising of 11,680 web-based sub-collections each. Additionally, wanting to conclude on the usefulness of such a clustering technique within the P2P IR setting, we looked into the distribution of relevant documents at various levels of cluster effectiveness.

By thoroughly evaluating our assumptions through experimentation we conclude that document ordering does not, practically, lead to different classifications. The main reason for this, perhaps unexpected, result is the failure of single-pass clustering to effectively discover clusters in the document collections. This lead to extremely low clustering ratios, i.e. in numerous very small clusters, across all document orderings and thresholds and in both testbeds we experimented with.

Perhaps this result mainly occurred because of the nature of web-based collections, having documents of variable length, with most being too small as well as of low quality. In that case, we can assert that the term-frequency vectors we used were not adequate to describe sufficiently well the underlying documents they represented. However, we are very sceptical against the thought that even changing the document descriptions would lead to significantly (not in the statistical meaning of the word) different, let alone, better classifications. Summarising, at least for web-based or other collections of similar properties to these we studied, single-pass clustering appears to be robustly inadequate.

Even though the results of this study seem to be rather conclusive, we do not believe that this is enough to dismiss the potential usefulness of single-pass clustering for distributed and P2P IR. Indeed, we are currently in the process of experimenting with alternative, better written and structured collections of documents such as these present in the collections of the AdHoc track of TREC. Another piece of research that we are planning to undertake is the comparison of the classifications we obtained against classifications that have been derived from the more theoretically justified agglomerative algorithms, such as Ward's method [2]. This would, at least, give us more indications for the usefulness of clustering within the context of distributed and P2P IR.

ACKNOWLEDGEMENT

We would like to thank Dr. Leif Azzopardi of the University of Strathclyde for his help during the initial stages of this work.

This work has being jointly funded by the EPSRC and Sharp Laboratories of Europe Ltd. (Project number: GR/P02653/01).

REFERENCES

[1] C. J. van Rijsbergen, *Information Retrieval*, 2nd ed. London: Butterworths, 1979.

[2] J. H. Ward, "Hierarchical grouping to optimize an objective function," *Journal of American Statistical Association*, vol. 58, no. 301, pp. 236–244, March 1963.

[3] F. Murtagh, "A survey of recent advances in hierarchical clustering algorithms," *The Computer Journal*, vol. 26, no. 4, pp. 354–359, 1983.

[4] Q. He, "A review of clustering algorithms as applied in ir," University of Illinois, Tech. Rep. UIUCLIS–1999/6+IRG, 1999.

[5] R. M. Cormack, "A review of classification," *Journal of the Royal Statistical Society*, vol. Series A, no. 134, pp. 321–353, 1971.

[6] I. A. Klampanos and J. M. Jose, "An architecture for information retrieval over semi-collaborating peer-to-peer networks," in *Proceedings of the 2004 ACM Symposium on Applied Computing*, vol. 2, Nicosia, Cyprus, March 14–17 2004, pp. 1078–1083.

[7] J. Lu and J. Callan, "Content-based retrieval in hybrid peer-to-peer networks," in *Proceedings of the twelfth international conference on Information and knowledge management*. ACM Press, 2003, pp. 199–206.

[8] B. Krishnamurthy, J. Wang, and Y. Xie, "Early measurements of a cluster-based architecture for p2p systems." San Francisco, USA: ACM SIGCOMM, November 2001, internet Measurement Workshop.

[9] M. Khambatti, K. Ryu, and P. Dasgupta, "Peer-to-peer communities: Formation and discovery." PDCS'02, November 2002.

[10] C. H. Ng and K. C. Sia, "Peer clustering and firework query model." Hawaii: 11th World Wide Web Conference, May 2002.

[11] I. A. Klampanos, V. Poznański, J. M. Jose, and P. Dickman, "A suite of testbeds for the realistic evaluation of peer-to-peer information retrieval systems." in *ECIR*, 2005, pp. 38–51.

[12] I. Soboroff, "Does wt10g look like the web?" in *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM Press, 2002, pp. 423–424.