A Suite of Testbeds for the Realistic Evaluation of Peer-to-Peer Information Retrieval Systems

Iraklis A. Klampanos[1], Victor Poznański[2], Joemon M. Jose[1],
and Peter Dickman[1]

[1] University of Glasgow, 17 Lilybank Gardens, G12 8QQ, Glasgow, U.K
{iraklis, jj, pd}@dcs.gla.ac.uk
[2] Sharp Labs of Europe Ltd., Edmund Halley Road,
Oxford Science Part, OX4 4GB, Oxford, U.K
vp@sharp.co.uk

Abstract. Peer-to-peer (P^{2}P) networking continuously gains popularity among computing science researchers. The problem of information retrieval (IR) over P^{2}P networks is being addressed by researchers attempting to provide valuable insight as well as solutions for its successful deployment. All published studies have, so far, been evaluated by simulation means, using well-known document collections (usually acquired from TREC). Researchers test their systems using divided collections whose documents have been previously distributed to a number of simulated peers. This practice leads to two problems: First, there is little justification in favour of the document distributions used by relevant studies and second, since different studies use different experimental testbeds, there is no common ground for comparing the solutions proposed. In this work, we contribute a number of different document testbeds for evaluating P^{2}P IR systems. Each of these has been deduced from TREC's WT10g collection and corresponds to different potential P^{2}P IR application scenarios. We analyse each methodology and testbed with respect to the document distributions achieved as well as to the location of relevant items within each setting. This work marks the beginning of an effort to provide more realistic evaluation environments for P^{2}P IR systems as well as to create a common ground for comparisons of existing and future architectures.

1 Introduction

Peer-to-Peer (P^{2}P) computing is a modern networking paradigm that allows for seamless communication of connected devices at the application level. In P^{2}P networks all participating processes are made equally capable, by exerting both server and client functionalities [1]. Because of this fact, and also because these networks are built on software, P^{2}P networking has become a fast-developing research field, since it can, potentially, provide cost-effective, efficient and robust solutions. Like in any distributed system, location and retrieval of relevant information and resources is of paramount importance. Therefore, depending on the application at hand, IR can be thought of as an important component of P^{2}P -based solutions. This follows from current P^{2}P applications (file-sharing, transparent interconnection of corporate sites etc. [2, 3]) as well as from potential uses (project collaboration, intelligent information sharing etc.). On the

D.E. Losada and J.M. Fernández-Luna (Eds.): ECIR 2005, LNCS 3408, pp. 38–51, 2005.

other hand, P^{2}P IR shares the aim of distributed IR, that is to achieve more effective IR than centralised solutions, through successful resource description, location and fusion of results[4].

P^{2}P IR networks have a number of inherent properties that render their evaluation a particularly hard task. First, they are usually assumed to be very large. Hundreds of thousands of computers participate typically in P^{2}P file-sharing networks. Researchers deal with such environments by simulating a carefully selected subset of their systems' intended functionality. Additionally, participating nodes are expected to join or leave unexpectedly and, moreover, nodes might leave willingly or simply crash, something which is easily resolvable in a medium-sized distributed system built on higher-end components. This effect is hard to simulate, however it is up to individual proposals to address how they deal with it.

On the IR side, in a P^{2}P network, the distribution of documents is, to a significant scale, a result of previous location and retrieval. However, this also depends on the application specification and/or on other non-functional requirements that may be imposed (such as copyright considerations etc.). Defining and simulating user behaviour, especially in a very large distributed system, is a complex and intimidating task. Indeed, most published P^{2}P IR solutions have dealt with this problem indirectly. Instead of simulating user behaviour, people have attempted to reflect it in the document distributions (or testbeds) they have used for their evaluation. The problem with such approaches is twofold. Firstly, there are cases where the distribution of documents does not reflect the application scenario successfully and therefore such evaluation results are hardly conclusive. Secondly, since individual proposals devise and use their individual testbeds, comparisons between different solutions, through their evaluation results, is impossible. Addressing these issues, we contribute a number of realistic testbeds, suitable for the evaluation of P^{2}P IR systems.

Emphasising on the fact that there may be many, diverse potential P^{2}P IR applications, we identify a number of possible scenarios and propose methodologies that can be used for the creation of realistic information-sharing testbeds. We have derived our testbeds using TREC's 10G Web collection (WT10g). This collection is an archive of 11,680 Web domains, 1.69 million documents and its relevance assessment comprises of 100 queries. This paper is organised as follows: The next Section is about related work of evaluating P^{2}P IR systems, which strengthens our reasoning in favour of the adoption of better thought-out evaluation environments. Section 3 presents a number of P^{2}P IR scenarios, their properties as well as a number of appropriate document testbeds that could address them. Section 4 presents an analysis of the obtained testbeds with respect to their document distributions among the derived peer-collections. Another aspect we have looked at is the distribution of the relevant, to the standard WT10G queries, documents. Finally, in Section 5, we present our conclusions regarding the current work and how this may relate to future P^{2}P IR systems.

2 Background and Motivation

The potential of P^{2}P architectures spans a number of possible applications. At the moment, the most popular ones are file-sharing (e.g. Limewire [5]) and distributed storage and retrieval systems (e.g. Freenet [6]). Future applications might include long-distance integrated development environments, virtual offices, P^{2}P photograph-sharing applications and other sophisticated information-sharing environments. This potential has been realised by the research community and there exist many ongoing projects that attempt to identify and solve related problems.

Properly evaluating research proposals, through simulation, as well as comparing, at least, systems that target at similar application domains, is a major part of research methodology. However, such provisions have not been taken for P^{2}P IR yet.

The research solutions that have been proposed to date, can be divided in two major fronts: Distributed-Hash-Table(DHT)-based [7, 8, 9] and content-based solutions [10, 11, 12, 13]. Although there has been some overlap between those two trends, the motivation behind them and, consequently, the solutions they propose focus on location and retrieval of different items of information. While a DHT is a convenient structure for location, routing and retrieval of items baring IDs or descriptions consisting of a few keywords, content-based approaches attempt to create informed networks by propagating knowledge and statistics about document collections. In many respects, DHT-based approaches fall within the realm of databases, while content-based approaches are, usually, IR-inspired. The work presented in this paper provides realistic testbeds for the effective evaluation of both types of systems. To the best of our knowledge, this constitutes the first attempt of its kind in the field of P^{2}P IR.

In this Section, we motivate our study by focusing on the evaluation of three content-based P^{2}P architectures. A summary of the evaluation characteristics of the following proposals can be found in Table 1. Although following different mechanisms, the main target of the following, cited, systems is to achieve effective resource selection and efficient routing, given a query. Success in achieving these goals translates directly into the retrieval being effective. We will not present the mechanisms these systems use in order to perform IR, since it is not within the intended focus of this paper. Instead we will be discussing their target application areas as well as the experimental testbeds in which they were evaluated.

2.1 SETS

SETS (Search Enhanced by Topics Segmentation) [11] is a P^{2}P IR system aimed at information-sharing (full-text search) over large, open P^{2}P networks. The idea behind SETS, and indeed other similar architectures, is to arrange peers in such a way that queries only have to traverse a small subset of the total participants in order to be effectively evaluated. SETS was evaluated in terms of query processing cost, that is the average number of sites that need to be contacted in order a query to be answered. The evaluation setup consisted of three different document sets, the TREC-1,2-AP, the Reuters collection as well as the CiteSeer database. Each site (peer) would hold documents of a particular author, therefore a very small number of mostly similar documents. In an open information-sharing network, however, such a setting could only occur during the network's bootstrapping phase. After that, users (authors in this paradigm), would be expected to search and download documents locally, which in turn would be searchable by others and so on. Therefore, in the end, the peer-collections would grow larger, the distribution of documents across the peers would start following more obvious power-law patterns, and replication of documents, through retrieval, would play a significant part in the effectiveness of any such $\mathrm{P}^2\mathrm{P}$ IR system.

2.2 A Hybrid, Content-Based $\mathrm{P}^2\mathrm{P}$ Network

Lu and Callan [12] proposed a hybrid $\mathrm{P}^2\mathrm{P}$ network for addressing the problem of location, query routing and retrieval in a digital libraries setting. The term "hybrid" is used to distinguish between unstructured $\mathrm{P}^2\mathrm{P}$ networks, where all nodes behave as equals in absolute terms, from structured ones, where there exists a division between administrative peers (directory nodes) and leaf peers. However, such separation of functionality does not imply a separation in capabilities. The network does not stop being a $\mathrm{P}^2\mathrm{P}$ one, since at any given time and possibly depending on the nodes' characteristics, any leaf node can become a directory one and vice versa. In this proposal, certain directory nodes were made responsible for holding indices of specific interest areas. In essence, the information providers were clustered according to content, and if they fell within more than one topic of interest, they were assigned to more than one directory nodes.

The authors evaluated their architecture by using TREC's WT10g collection. For these experiments, 2,500 collections (domains) were randomly selected, containing 1,421,088 documents, and then they were clustered. The algorithm used was a soft-clustering [14] one, so that collections that were about more than one topic, got assigned to more than one clusters. By clustering, the authors managed to simulate the organisation of similar topics around their corresponding directory nodes in the network. The measurements taken were precision, recall and the number of messages generated for each query. Even though this testbed is suitable for a digital library scenario, it would be interesting to be able to evaluate this system in different settings, that exhibit different document distributions. Furthermore, the use of clustering might have enforced a more rigid organisation of content than the one observed in real-life digital library scenarios. However, recognising the importance of this testbed, we have included and analysed it in our study too.

2.3 IR in Semi-collaborating $\mathrm{P}^2\mathrm{P}$ Networks

Same as the above, this system [13] is also a hybrid one. The intended target domain is large, information-sharing networks. The term "semi-collaborating" implies that, although peers do not need to share internal (and possibly proprietary) information, they do need to share information about their shared document collections. The testbed used for the evaluation of this architecture was based on the TREC adhoc collection, comprised of 556,077 documents. Also, the relevance assessments from TREC 6 and 7 were used, featuring 100 queries. The number of peers simulated was 1,500. Because of difficulties to cluster the whole collection using agglomerative approaches, the authors distributed the relevant documents of the topics to a small number of peers. The rest of the documents were assigned randomly to the peer population. Admittedly, this evaluation strategy has a number of serious drawbacks. Firstly, distributing the great majority of the documents randomly to peer-collections, is something unrealistic in an information-sharing scenario. On the other hand by assigning the relevant documents of the queries to some peers, and then by evaluating the system using the same queries, can produce results that are inconclusive and can even be considered as erroneous.

Table 1. An overview of the evaluation environments of three sample P^{2}P IR proposals. The incompatibilities are evident

Architecture	*Collection(s)*	*Number(s) of Peers*	*Avg. Num. of Documents*
SETS [11]	TREC AP / Reuters / Citeseer	1,834 / 2,368 / 83,947	43 / 44 / 5
HYBRID [12]	TREC WT10g	2,500	568
S-C P^{2}P IR [13]	TREC Adhoc	1,500	370

2.4 Summary

The evaluation of P^{2}P IR architectures is a complex task, which is usually done through simulation. However, in many proposed systems, the evaluation testbeds used only reflected a very small subset of possible application scenarios, and sometimes, even unrealistic ones. Hence, some of the results of such evaluations can be thought to be inconclusive. Additionally, the diversity of the evaluation testbeds used in different studies, prohibit the fruitful comparison between, even, systems that aim at similar information environments. For example, it would be interesting to compare the systems presented above in different experimental settings, as this would reveal their strengths and weaknesses at different information-sharing environments. We address these issues by providing a number of realistic testbeds for the evaluation of P^{2}P IR architectures. We reason in favour of our testbeds' appropriateness based on both the methodology used to derive them (described in the next section) as well as on their document distributions and other properties (presented in Section 4).

3 Testbeds for P^{2}P IR

By studying existing P^{2}P networks and various proposed solutions, such as the ones discussed in the previous section, we have identified a number of different features that could potentially affect IR. In this section, we present three high-level scenarios that should exhibit different characteristics, along with suitable testbeds that could be used for P^{2}P IR evaluation.

In P^{2}P information sharing networks, like in other distributed IR systems (such as the Web, digital libraries or P^{2}P file-sharing) each participating node shares documents about a limited number of topics. In other words, it is rather unlikely that random content will be placed into any node of such networks. Moreover, it has been shown that in file-sharing P^{2}P networks, files are distributed in power-law patterns across participating peers[15]. Therefore, these properties should be preserved in realistic P^{2}P IR testbeds as well.

Another important aspect of information-sharing environments is content replication. It has been shown by various studies that replication can affect retrieval and that it is even a desirable feature in some cases [16, 17]. Typically, replication occurs as a result of previous querying and retrieval. However, there are cases where retrieving content freely cannot be allowed because of either copyright issues or ethical considerations etc. An example of such a case could be a P^{2}P photograph-sharing application, where people might want to share their photographs with a limited number of people, family or others, at the time of their choosing, without compromising their privacy. Therefore, we feel that suitable testbeds for P^2P IR architectures should address both situations. Following from that, each of our testbeds comes in two flavours, one with included replication and one without. In this work, the names of the testbeds with replication have been suffixed by *WR*, while those without replication have been suffixed by *WOR*. The testbeds presented in this study can be reproduced by downloading the corresponding definitions from *http://www.dcs.gla.ac.uk/∼iraklis/evaluation*.

3.1 Information-Sharing Environments

Currently, information-sharing scenarios are the most popular ones. They reflect settings analogous to the widely used file-sharing P^2P networks like Gnutella [2]. In such settings, the document distribution among the participating peers follows power law patterns [15]. The same is true for the world-wide Web, where there is a power-law distribution of documents within Web domains. In order to address this fact, we chose to represent each peer collection by one Web domain. By following this simple procedure we both get a power-law distribution of the documents in the network and also a large enough number of peers to drive potential simulations (11,680 for TREC's WT10g collection).

A replication effect can be achieved, if desired, by pulling into a peer-collection all other documents, residing at different domains, pointed by the documents of the current domain. In other words we exploit inter-domain links between Web domains in order to achieve meaningful replication in our P^2P IR testbed. The intuition behind this is straightforward: if a Web site links to another external Web page, these must be related in some way. Therefore, it would make more sense to replicate as described than to pull documents randomly into the peer-collections.

This set of testbeds was derived by using the Web domains unchanged and so it was named *ASIS*. Therefore, by following the naming convention described above, for the ASIS case, we have two testbeds: one with replication – *ASISWR* – and one without – *ASISWOR*.

3.2 Uniformly Distributed Information Environments

This testbed can be used for the simulation of systems where the documents are distributed uniformly across the peer population. Such distribution could result from limited I/O capabilities or memory of the participating devices, copyright issues or in the case of simulating IR behaviour in loosely controlled grid networks.

This testbed was obtained by dividing the available web domains into three buckets – under-sized, over-sized and properly-sized – according to the number of documents they share. Then, we moved each excessive document from the over-sized bucket into its closest under-sized domain; closeness defined as the cosine similarity between the page to be moved and the homepage of each of the under-sized domains. Once an under-sized domain or an over-sized domain reached the desired number of documents, they were moved into the properly-sized bucket.

We chose to use homepages because of efficiency reasons as well as because of the fact that homepages are written to be found and read and should, therefore, describe, to some extent, the rest of the Web-site. Some of them do that successfully and others do not; similarly, in a P^2P network we would expect some peers to share content consistently about a number of topics, in contrast to other peers. Using homepages provides us with an intuitive parallelism between Web-sites and peer-collections.

Like the ASIS testbed, this too has two versions: one with replication – *UWR* – and one without – *UWOR*. The replication method used is the same as in the ASIS testbed.

3.3 Digital Libraries

P^2P IR solutions could also aid the effective organisation and retrieval in distributed digital library (DL) environments. This fact has also been addressed in [12], as mentioned in Section 2. In a digital library setting we would, typically, expect to have fewer remote collections than in the other settings described above. However, we would also expect individual libraries to hold more documents, on average, than peer-collections would in an information-sharing scenario. The distribution of documents, therefore, would be expected to follow a power-law pattern, although perhaps not as an extreme one as in an open information-sharing environment.

In order to obtain this testbed, we first selected the 1,500 largest domains. Then, we pulled each one of the remaining domains to the closest of the larger ones. Again, closeness was computed as the cosine similarity between the homepages of the related domains.

Similarly to the testbeds described above, this also comes in one version with replication –*DLWR* – and one without – *DLWOR*.

As mentioned above, our digital library family of testbeds also includes the one generated and used by Lu and Callan in [12], herein referred to as *DLLC*.

4 Analysis and Results

In this Section, we analyse the six testbeds previously created as well as the one used by Lu and Callan [12] (*DLLC*). Our intention is to provide insight and justification for the usefulness of these testbeds, not to make comparisons between any two of them. We first present and discuss the document distributions that we obtained from the various testbeds (Section 4.1). Then, we look at the distributions of relevant documents within the testbeds from two perspectives. In Section 4.2 we look at the number of collections needed to reach 100% recall (for the topics used), while in Section 4.3 we look at precision levels for the same level of recall.

Some of the general properties of these testbeds are summarised in Table 2.

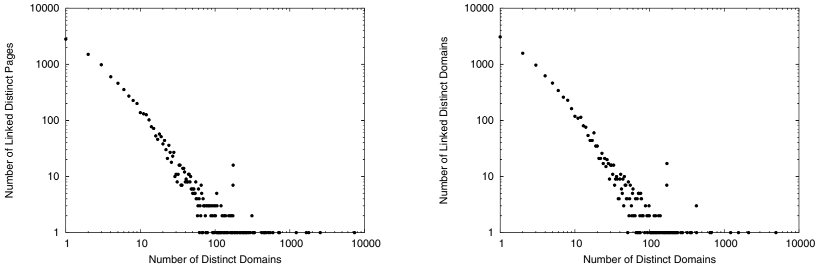
4.1 Document Distributions of the Testbeds

The distribution of the documents in a testbed reflects different possible scenarios, and can indeed affect the effectiveness of retrieval. While creating our testbeds, we took document distributions under consideration.

We already know that there is a power-law distribution of documents within the domains used in WT10g [18]. Therefore, exactly the same distribution of documents holds for the ASISWOR testbed. The imposition of replication via the method described

Table 2. General Properties

Testbed	*Num. of Collections*	*Num. of Documents*
ASISWOR	11,680	1,692,096
ASISWR	11,680	1,788,248
UWOR	11,680	1,692,096
UWR	11,680	1,788,041
DLWOR	1,500	1,692,096
DLWR	1,500	1,740,385
DLLC	2,500	1,421,088



(a) Domain-to-page links distribution

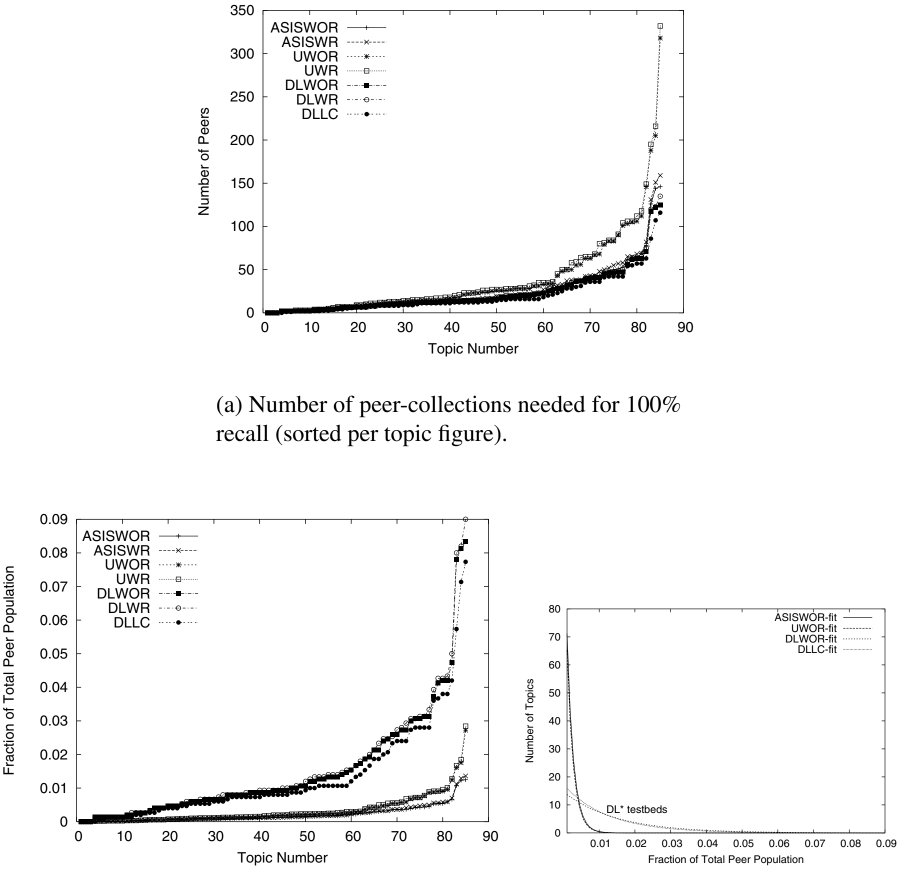
(b) Domain-to-domain links distribution

Fig. 1. The distribution of inter-domain links in WT10g

in 3.1 did not alter this distribution, since the distribution of outgoing domain-to-page inter-domain links is also a power-law one (Fig. 1(a)), just like the domain-to-domain distribution of links is (Figure 1(b)).

Uniformity was imposed on WT10g in order to obtain the UWOR testbed (Section 3.2). Because of the distribution of inter-domain links, however, UWR has lost this uniformity, even though on a small scale. We consider this effect to be adding to the testbed being realistic. We would expect that the document distribution of any initially uniformly distributed network would start skewing, over time, towards power-law patterns. That would happen if free replication was allowed at some point during the lifetime of the network.

Finally, the digital-library testbeds (DLWOR, DLWR and DLLC), also exhibit power-law document distributions. For DLWOR and DLWR testbeds (Section 3.3), the initial largest domain exhibit power-law document distributions. The further agglomeration of the rest of the domains only adds to the asymmetry of the distribution. The reason behind this effect is that the larger domains are bound to be attached to smaller ones since they usually cover a broader range of topics, which are also usually reflected in their homepages. A homepage of a portal, like *Yahoo!* for instance, will typically contain keywords relevant to a very large number of topics. For DLLC, although a soft clustering algorithm was used, the same reasoning should hold.



(a) Number of peer-collections needed for 100% recall (sorted per topic figure).

(b) Fraction of the peer population needed for 100% recall (sorted per topic figure).

(c) The distributions of the relevant documents in the testbeds (Exponential fits of the data).

Fig. 2. The distribution of relevant documents

4.2 The Location of the Relevant Documents (Recall)

In the second stage of our analysis we investigate the location of relevant documents in the testbeds. In particular, we are interested in the number of peer-collections that a query would have to be forwarded to, in order to obtain 100% recall. In other words we need to know how the relevant documents get distributed in the testbeds, for any one topic. Such information may be important to P^{2}P IR architectures that might want to exploit it in their resource selection and routing algorithms.

Figure 2 shows the distribution of the relevant documents in the various testbeds. In Fig. 2(a), we have sorted the topics according to the number of peer-collections that contain at least one relevant document. It can be seen that in the uniformly distributed testbeds (UWOR and UWR), a significantly larger number of peers need to be reached in order to achieve 100% recall. The rest of the testbeds need a similar number of collection in order to reach the same amount of recall. This might be expected since the uniformity imposed on the UWOR and UWR testbeds means that each collection shares a relatively small number of documents, so there is a higher probability that the relevant documents for some topics will be scattered among a larger number of collections.

In Fig. 2(b) we see the same information, this time presented against the fraction of the total peer population that each topic needs to reach to satisfy 100% recall. From this perspective it can be seen that the DL testbeds need to reach a higher fraction of the population, while the U* and ASIS* testbeds need a significantly lower fraction. This can be explained by the fact that in the DL testbeds we have created a much smaller number of peer-collections (1,500 and 2,500) than in the rest (11,680). This is also reflected in Fig. 2(c), where we have plotted the exponential fits for the *WOR distributions as well as for DLLC. The *WR testbeds follow similar distributions hence they were omitted.

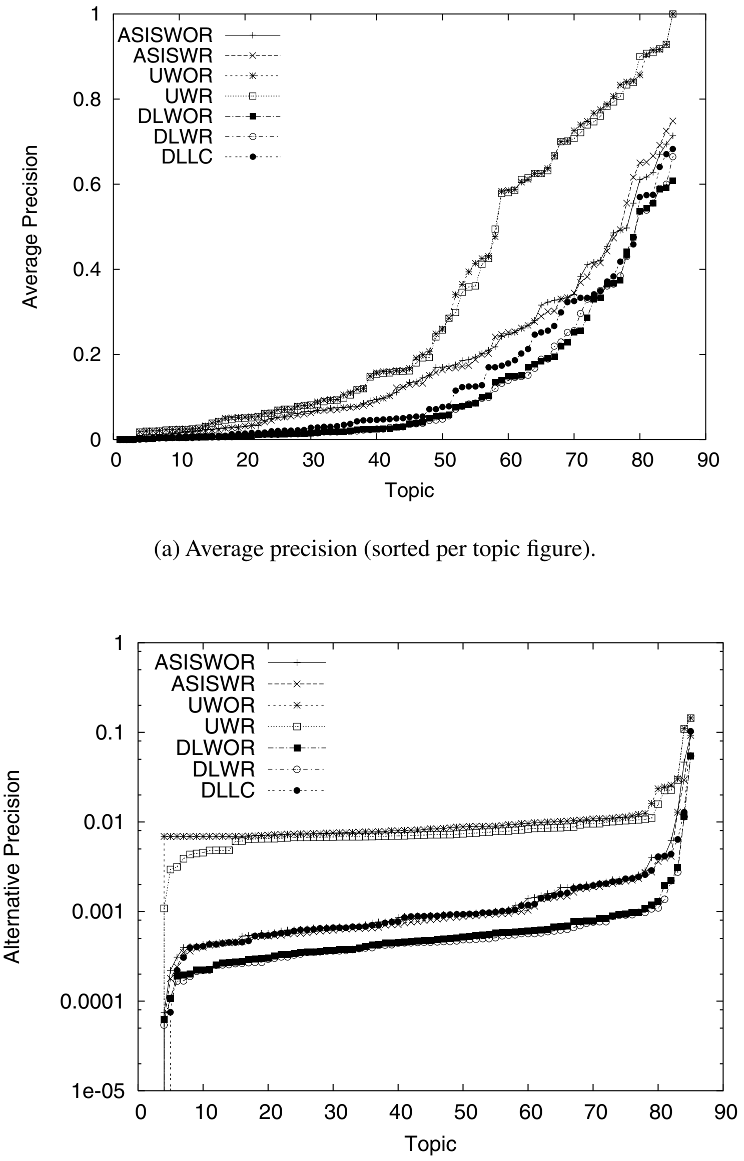
4.3 Coverage of the Topics (Precision)

Following the creation of the testbeds, another major aspect we looked at was the proportion at which topics were represented within the peer collections; in other words, the precision within the peer-collections. We have looked at precision from two different viewpoints[1]. Firstly, for each topic, we considered all the peer collections that had at least one relevant document and measured their average precision, *i.e.* $P_{\text{avg}} = 1/n \sum_{i=1}^{n} P_i$, where n is the number of peer-collections that have at least one relevant document and P_i is the precision as measured by the number of relevant documents over the total number of documents shared in the ith collection. These measurements are depicted in Fig. 3(a). Another way to look at precision was to consider the same peer-collections as one and then measure precision, *i.e.* $P_{\text{alt}} = \sum_{i=1}^{n} r_i / \sum_{i=1}^{n} total_i$, where r_i is the number of relevant documents of the ith collection and $total_i$ is the total number of documents shared by the ith collection. The alternative precision measurements are shown in Fig. 3(b).

The average precision measurements appear to be quite promising as to what a well designed P^{2}P IR architecture can potentially achieve. Although approximately half of the topics appear to be represented at a level of precision lower than 0.2, the rest follow an exponential increase, which reaches even 1 for one topic in the uniform testbeds. Overall, the uniformly distributed testbeds appear to perform a lot better, in terms of average precision, than the other ones. This can be explained by the fact that their collections share a small number of documents without great deviations. The second best-performing set of testbeds are the ASIS ones, and this is probably because of the cohesion that some domains demonstrate as to the topics they address. Finally, the DL* testbeds follow, whose collections share a larger number of documents.

In Fig. 4.3(b) we present the alternative definition of precision for the testbeds generated. The y-axis is presented in logarithmic scale for increased readability. Again the uniform distributions appear to be exhibiting higher levels of precision, although extremely lower than previously. The ASIS testbeds follow approximately the DLLC

[1] By the term “precision” we mean the coverage of specific topics in peer-collections. We do not imply that any actual retrieval took place.



(a) Average precision (sorted per topic figure).

(b) Alternative precision (sorted per topic figure).

Fig. 3. Precision in collections that achieve 100% recall

testbed, while the DLWOR and DLWR appear to be the worst in that respect. It is interesting to note the reason why DLLC appears to have higher precision levels than DLWOR and DLWR. Two explanations can be given for this artifact: firstly, DLLC's collections share less documents in total, and secondly, DLLC was generated by applying a soft-clustering algorithm. Therefore, DLLC should have a better concentration of relevant content within its collection than the other DL testbeds.

4.4 Discussion and Summary

All testbeds described share a number of features. Any of the topics included, needs to reach only a small number of collections in order to be fully met. This fact clearly stresses the need for well informed networks that exhibit effective resource selection and routing. Additionally, a large number of irrelevant documents are bound to reside at the same peers, therefore impeding the local retrieval systems as well as fusion.

Analysing the coverage of all the relevant collections as a single one has the following significance. A system that wants to achieve 100% recall, will have to reach all these collections, for a given topic. At the end of a session, a significant number of results might be returned to the initiator of a query, which will then have to fuse them, before presenting them to the user. Both the large number of peers that will be returning responses as well as the fraction of relevant over the total number of responses can seriously impede effectiveness. Based on the results presented in this study, we believe that, regardless of the retrieval mechanisms used at the peers, the lack of a highly effective fusion technique will have a very negative impact on any P^2P IR application, especially those that require high precision and lower recall.

Summarising, we would like to emphasise on the scenarios targeted by our testbeds and reason towards their usefulness. The ASIS* testbeds are targeted on simulating openly available information-sharing P^2P networks, *i.e.* potential networks and applications where users can retrieve, download and replicate other documents, as well as introduce their own. The reasoning behind this assertion is that, the ASIS testbeds exhibit power-law document distributions, that are found in file-sharing P^2P networks, the Web and elsewhere. Additionally, since we have used the Web domains unchanged, the documents in the deriving collections are bound to be loosely organised on content, *i.e.* they are not randomly allocated. Thirdly, the addition of replication, in the ASISWR testbed, addresses a potentially significant side-effect of information-sharing networks. Lastly, by using the Web domains, we achieve to obtain a relatively large number of peer collections, suitable for adequate evaluation.

The uniform testbeds are suitable for evaluating P^2P IR systems targeted at a different class of information-sharing environments. Such possible application include grid-like environments, where an equal amount of load is imposed on all participating nodes. Other possible scenarios include systems whose peers have limited I/O and memory capabilities (for example mobile devices), and therefore the addition of large numbers of new documents is impossible. Another relevant situation would be where replication through retrieval is not permitted because of various non-functional requirements. We believe that the U* testbeds are suitable for the evaluation of such systems because they incorporate a sufficiently large number of peers, the documents are uniformly distributed across the peer population, but still the documents shared by any peers are loosely related without, however, having been properly clustered.

Finally, the DL* testbeds would be suitable for a number of digital-library instantiations of the P^2P IR problem. These might include P^2P networks that bridge corporate information sites, Internet meta-searching, academic P^{2}P networks etc. The document distribution in these testbeds follows power-law patterns, as one would expect in the aforementioned scenarios, but the average number of documents shared is significantly higher than the other testbeds. Content consistency has been preserved by having each peer-collection represented by a number of loosely related Web domains.

5 Conclusions and Future Work

Evaluating IR architectures and systems for P^{2}P networks is a demanding and neglected task. In this paper we address the importance of using realistic document testbeds for the evaluation of P^{2}P IR architectures, something which has been overlooked by many studies published so far. For this reason, we provide a number of realistic testbeds addressing different application scenarios (summarised in Section 4.4). These testbeds are derived from the TREC WT10g collection, by following different methods of distributing its documents into a sufficiently large number of smaller peer-collections. Subsequently, we analyse our testbeds from a number of different perspectives in order to understand their properties as well as to obtain justified hints on what would be needed by any architecture in order to provide effective and efficient IR over a P^{2}P network.

From our analysis we draw the following conclusions. Firstly, fusion needs to be seriously looked at if we want to achieve high effectiveness and user satisfaction in future P^{2}P IR systems. Additionally, the fact that only a small proportion of the total peer-population suffices in order to achieve high recall, is a promising fact with respect to the efficiency of these networks. On the other hand, in order for a system to be able to identify and properly use the resources available, a lot of effort will have to be put both into the content-based organisation of the network as well as into its resource selection and query routing algorithms. Even though these needs have been addressed repeatedly in the literature, we have managed to observe them in a number of different evaluation settings.

There are many ways in which this work can be used and extended. A first step would be to use the testbeds for evaluating existing or newly proposed P^{2}P IR architectures in order to observe how their effectiveness changes in different environments. The adoption of a set of standard testbeds could provide a strong lead towards benchmarking studies for P^{2}P IR systems. Additionally, we could start looking at some temporal properties of P^{2}P networks and their effect on IR. Such properties might be the generation and growth of the network as well as the joining and leaving of nodes. Finally, we could use these testbeds in order to derive a series of stress tests for potential systems. The dynamics of P^{2}P networks is an area still under heavy research and exploration, without mentioning the effects it might impose on IR.

References

1. Oram, A., ed.: PEER-TO-PEER: Harnessing the Power of Disruptive Technologies. O'Reilly & Associates, Inc., CA 95472, USA (2001)
2. OSBM LLC.: The homepage of gnutella. http://www.gnutella.org/ (2003)

3. Groove Networks: The homepage of groove networks. (http://www.groove.net/) As viewed on March 27 2004.
4. Callan, J.: 5 – Distributed Information Retrieval. In: Advances in Information Retrieval. Kluwer Academic Publishers (2000) 127–150
5. Lime Wire LLC.: The homepage of limewire. http://www.limewire.com/ (2003)
6. Clark, I.: The homepage of freenet project. http://www.freenet.sourceforge.org/ (2003)
7. Rowstron, A., Druschel, P.: Pastry: Scalable, decentralized object location, and routing for large-scale peer-to-peer systems. Lecture Notes in Computer Science **2218** (2001)
8. Hildrum, K., Kubiatowicz, J.D., Rao, S., Zhao, B.Y.: Distributed object location in a dynamic network. In: Proceedings of the Fourteenth ACM Symposium on Parallel Algorithms and Architectures. (2002) 41–52
9. Ratnasamy, S., Francis, P., Handley, M., Karp, R., Shenker, S.: A scalable content addressable network. In: Proceedings of ACM SIGCOMM 2001. (2001)
10. Cuenca-Acuna, F.M., Peery, C., Martin, R.P., Nguyen, T.D.: PlanetP: Using Gossiping to Build Content Addressable Peer-to-Peer Information Sharing Communities. In: Twelfth IEEE International Symposium on High Performance Distributed Computing (HPDC-12), IEEE Press (2003)
11. Bawa, M., Manku, G.S., Raghavan, P.: Sets: search enhanced by topic segmentation. In: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval, ACM Press (2003) 306–313
12. Lu, J., Callan, J.: Content-based retrieval in hybrid peer-to-peer networks. In: Proceedings of the twelfth international conference on Information and knowledge management, ACM Press (2003) 199–206
13. Klampanos, I.A., Jose, J.M.: An architecture for information retrieval over semi-collaborating peer-to-peer networks. In: Proceedings of the 2004 ACM Symposium on Applied Computing. Volume 2., Nicosia, Cyprus (2004) 1078–1083
14. Lin, K., Kondadadi, R.: A similarity-based soft clustering algorithm for documents. In: Proceedings of the 7th International Conference on Database Systems for Advanced Applications, IEEE Computer Society (2001) 40–47
15. Saroiu, S., Gummadi, P.K., Gribble, S.D.: A measurement study of peer-to-peer file sharing systems. In: Proceedings of Multimedia Computing and Networking 2002 (MMCN '02), San Jose, CA, USA (2002)
16. Lv, Q., Cao, P., Cohen, E., Li, K., Shenker, S.: Search and replication in unstructured peer-to-peer networks. In: ICS, New York, USA (2002)
17. Cuenca-Acuna, F.M., Martin, R.P., Nguyen, T.D.: Planetp: Using gossiping and random replication to support reliable peer-to-peer content search and retrieval. Technical Report DCS-TR-494, Department of Computer Science, Rutgers University (2002)
18. Soboroff, I.: 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) 423–424