Evaluating Document Retrieval Methods for Resource Selection in Clustered P2P IR

Rami S. Alkhawaldeh
The University of Glasgow
University Avenue, G12 8QQ
Glasgow, United Kingdom
r.alkhawaldeh85@gmail.com

Joemon M. Jose
The University of Glasgow
University Avenue, G12 8QQ
Glasgow, United Kingdom
Joemon.Jose@glasgow.ac.uk

Deepak P
Queen's University
University Road, BT7 1NN
Belfast, United Kingdom
deepakp7@gmail.com

ABSTRACT

Resource Selection (or Query Routing) is an important step in P2P IR. Though analogous to document retrieval in the sense of choosing a relevant subset of resources, resource selection methods have evolved independently from those for document retrieval. Among the reasons for such divergence is that document retrieval targets scenarios where underlying resources are semantically homogeneous, whereas peers would manage diverse content. We observe that semantic heterogeneity is mitigated in the clustered 2-tier P2P IR architecture resource selection layer by way of usage of clustering, and posit that this necessitates a re-look at the applicability of document retrieval methods for resource selection within such a framework. This paper empirically benchmarks document retrieval models against the state-of-the-art resource selection models for the problem of resource selection in the clustered P2P IR architecture, using classical IR evaluation metrics. Our benchmarking study illustrates that document retrieval models significantly outperform other methods for the task of resource selection in the clustered P2P IR architecture. This indicates that clustered P2P IR framework can exploit advancements in document retrieval methods to deliver corresponding improvements in resource selection, indicating potential convergence of these fields for the clustered P2P IR architecture.

Categories and Subject Descriptors

H.4.2 [**Information Systems Applications**]: P2P Information Retrieval

Keywords

Adapted resource selections; Information Retrieval; Peer-to-Peer; Clustering peers; Content-based; Evaluation.

1. INTRODUCTION

Federated search provides a uniform interface across plurality of searchable resources by way of a broker. P2P IR systems do federated search over multiple peers, each of which manages a subset of the full dataset. In cooperative P2P IR environments, the brokers acquire content information from their resources beforehand and could potentially build indexes upon such information; these are then used to make a decision regarding routing a given query to those peers that are most likely to contain relevant documents [5, 8]. This contrasts with un-cooperative environments, where the resources provide only the search API. Resource selection (i.e., Query routing) is the problem of selecting a subset of relevant peers for a given query in P2P IR systems [18]. Following resource selection, the query is sent to the selected resources which would then process it and send results back to the broker. The broker would merge results from across peers providing a single ranked list of results to the user. Resource selection is a critical component in P2P IR; low-quality resource selection, the case where the relevant peers get excluded would inevitably lead to less effective IR results. In this paper, we consider the applicability of document retrieval methods for resource selection in a well-studied co-operative P2P IR environment that uses text clustering to structure content within peers [10].

There has been a large amount of work on resource selection mechanisms for general co-operative P2P IR where the broker operates on a set of resources, each of which manage a subset of the dataset. In a sense, the problem of resource selection may be seen as analogous to the problem of selection of relevant documents in conventional IR systems. In a federated P2P IR based scholarly article search system, the allocation of documents to resources could be dependent on factors such as the publication venue and publisher, which are typically completely orthogonal to the content in the documents. Thus, each resource could comprise documents that are semantically as diverse as the corpus itself. This diversity of topics at the level of each resource breaks the analogy between resource selection and document retrieval, since documents are typically focused on topic or a small number of topics and are thus limited in their heterogeneity. This, among other factors, has led to divergence between the techniques for resource selection and document retrieval, with each stream evolving independently. For example, popular resource selection methods such as CVV [20] and Taily [2] use information about word distributions across peers, whereas popular document retrieval models rely more on simpler term weighting (e.g., tf-idf) and divergence from randomness (e.g., BB2 [3]).

There has been renewed interest in the clustered P2P IR architecture [10, 1, 13] that employs text clustering to build a two-tier structure. The usage of text clustering in the architecture ensues that routing decisions be made at the level of resource groups that are internally homogeneous. This property of the clustered P2P IR architecture also brings resources conceptually closer to documents in the sense of not

CIKM '16, October 24–28, 2016, Indianpolis, IN, USA.

DOI: http://dx.doi.org/XXXX.XXXX

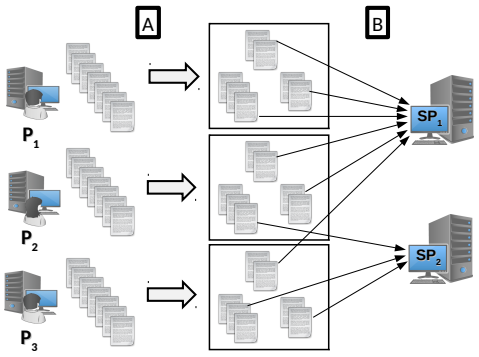

Figure 1: Clustered 2-Tier Architecture

being very divergent topically. In this paper, we posit that the clustered P2P IR architecture necessitates a re-look at the applicability of document retrieval methods for resource selection. Our main contributions are as follows:

- We consider the applicability of document retrieval methods for resource selection in the context of clustered P2P IR where routing decisions are to be made among semantically coherent resources.
- We do an empirical benchmarking of document retrieval methods against state-of-the-art resource selection methods, on the clustered P2P IR architecture. Our results across many testbeds establish that document retrieval methods are able to deliver consistently superior accuracies in resource selection.

We start by describing the clustered P2P architecture, our target environment, and outline reasons as to why we think that would be a friendly environment for document retrieval methods. We then summarize the document retrieval and resource selection methods that we benchmark in this study, followed by our empirical analysis and conclusions.

2. CLUSTERED 2-TIER ARCHITECTURE

Figure 1 illustrates the construction of the clustered 2-tier architecture. Each of the peers maintain a subset of documents, as shown by the different P_is in the left side of the figure. The subset of documents within each peer are subjected to a clustering process, illustrated in the figure as Step **A**; we will call this as *intra-peer clustering*. Though the figure shows 3 clusters consistently for every peer, there could in general be any number of clusters. Phase **B** clusters these intra-peer clusters, across peers, into a specified number (two, in the figure) of clusters. Each such cluster is managed by a super-peer (SP_i). Due to the clustering, not every super-peer necessarily would have representation from each peer; in our example, SP_2 does not have representation from P_1. The super-peer level, as may be noted, is an additional layer giving the framework the name 2-tier. Every query to the P2P IR system is sent to *each* of the super-peers, which would then use the information from the intra-peer clusters it manages, to route the query to one or more peers to which it is connected.

For a news search engine where different peers manage content from separate news agencies, sports related news articles may be separated out from others within each peer in the intra-peer clustering phase. Due to the second-level clustering, the sports clusters from separate peers are expected to be collected into a super-peer. Thus, the 2-stage clustering process ensures that routing decisions can be made at the level of super-peers that manage coherent content internally as well as among each other, while not disturbing the document assignment to peers; this likens the scenario to a domain-specific search at each super-peer.

Motivation: In general P2P IR without any intra-peer clustering, different resources at the resource selection level could potentially have widely varying content; this makes it infeasible to consider resources as analogous to documents, since document processing techniques often are built with the assumption that only a few topics are touched upon, in each document. An example is the concentration parameter in LDA that enforces topic sparsity at the document level. As another example, two large documents comprising text segments from across domains could achieve a high tf-idf cosine value due to a lot of small similarities adding up and end up being competitive with a pair of documents from the same domain; thus, tf-idf cosine is better applicable in cases where documents exhibit a good amount of lexical skew (akin to topical focus). In short, the absence of content coherence in general P2P IR architectures has made it infeasible to exploit the advances in document processing directly for resource selection, resulting in a divergent evolution of techniques for the tasks, as discussed earlier. Our central hypothesis is that the clustered P2P architecture, by virtue of clustering, helps bring back the analogy between documents and resources, thus recalling document relevance methods into contention for resource selection.

Type	Technique	Remarks
Vector Space	TF-IDF	TF-IDF Cosine similarity b/w Query and Document Vectors
Document Relevance	Okapi.BM25	Combination of binary independence model with term frequency saturation.
Query Relevance	LM	Language Model based assessment of query relevance
Divergence from Randomness	BB2 In_expB2 In_expC2 InL2	Bose-Einstein Model Inv Exp. Doc. Freq. Model Inv Exp. Doc. Freq. Model Inv Doc. Freq. Model

Table 1: Document Retrieval Models

3. DOCUMENT RETRIEVAL METHODS

Table 2 summarizes the document retrieval methods that we will use in our benchmarking study. The vector space model-based tf-idf scheme [4] quantifies the cosine similarity between tf-idf vectors, without treating the query and the documents differently. We also employ probabilistic retrieval models such as Okapi.BM25 [15] and Language Modeling (LM [9]) in our analysis. These differ in terms of the conceptualization of the model they create; BM25 models relevant documents given a query, whereas LM models query relevance using the document as a reference. The third type of models quantify divergence from randomness, based on the idea that the informativeness of a term may be measured by examining how much the term frequency distribution departs from a benchmark distribution generated by a random process. The different techniques in this family differ in terms of the randomness model they employ (listed in the table), and the formulation used in the after-effect and normalization steps. The details of these divergence models appear in [3]; a webpage[1] summarizes them succinctly.

Usage for Resource Selection: The above techniques, with the exception of TF-IDF (e.g., [14] aggregates TF-IDF across documents in a peer), have not been exploited much

[1] http://terrier.org/docs/v3.5/dfr_description.html

Method	Resource Score Computation
CVV	$\propto \sum_q CVV(q) \times DF_P(q)$ where $CVV(q)$ is the variance of q across peers and $DF_P(q)$ is the total frequency of q in the peer P
CORI	$\propto \sum_q \frac{DF_P(q)}{DF_P(q)+\alpha+\beta\times\#words\_in\_P}$, $DF_P(Q)$ defined as for CVV
vGlOSS	Estimate of #docs the peer would return assuming query words co-occur in docs
KL	LDA-based topic-wise PDFs over words from each peer compared against topic dist from Q using KL-divergence
Taily	Approximates distribution of document scores in each resource using a gamma dist. and scores resources based on the estimate of high-scored documents from the distribution

Table 2: Resource Selection Models for Co-operative P2P IR (Q stands for query)

for resource selection in hierarchical P2P IR, such as our setting. Since these techniques are specialized to ranking documents, i.e., sets of words, resources under each super-peer need to be modelled as documents. We do this by simply using the *big document* model [17]. Under this model, one big document is created for each resource managed by the super-peer; as a toy example, $SP2$ in Figure 1 would be searching over two documents, one built by collating the documents it manages from P3 and the other formed by the document subset of P2. Having defined the big documents, resource selection is just about ranking the big documents using the chosen model and routing the query to the resources corresponding to the top-ranked big documents.

4. CO-OPERATIVE P2P IR RESOURCE SELECTION METHODS

We now summarize the co-operative P2P IR resource selection methods used in our study, in Table 3. CVV [20] scores resources by preferring those that have a high concentration of unevenly distributed (across resources) query terms; it may be noted that in scenarios such as our news search example, term occurrences are expected to not vary much across resources. CORI [5] generalizes Okapi.BM25 to arrive at a resource level probabilistic scoring. vGlOSS [8] scores resources using an estimate of the number of relevant documents in them making use of a query word co-occurrence assumption. KL [19] makes use of topic distributions learnt using intra-peer clustering, and scores resources using the KL-divergence of the query topic distribution from that of each resource. Taily [2], the most recent technique, fits a gamma distribution over the scores for documents (wrt the query) within each resource. This, along with the size of the resource, is used to estimate the number of top-scored documents for each query within the resource. We do not use the decision-theoretic framework [7] in our analysis since it requires (as noted in [16]) extra information such as relevance judgements not presumed in other methods.

Usage in Clustered P2P IR: Each of these resource selection methods are used at each super-peer, the resource selection point, in the clustered 2-tier architecture.

5. EVALUATION FRAMEWORK

For a given testbed, comprising a document corpus split into peer-level collections (with or without replication), we first do the two-level clustering to instantiate the clustered 2-tier architecture. We use repeated bi-sect k-means for intra-peer clustering upto a cluster size threshold of 5. This is followed by a K-means clustering in the second phase with $k = 50$ so that the intra-peer clusters are distributed among 50 super-peers [1]. Our query set is a set of 100 simulated queries generated from TREC topics 451-550[2]. Having setup the framework, we use each method chosen for the study to do resource selection on each super-peer in response to each query; for a fixed choice of the selectivity parameter (say, 5%), that much of peers are selected within each super-peer. The query is then sent to the chosen resources (across super-peers) who in turn perform document retrieval using the Okapi.BM25 model; their results are collected and merged using the well-known COMBMNZ merging strategy [12]. The final merged result set is evaluated against classical IR parameters such as precision and MAP, as well as recall since labelled relevance information is available for the query-testbed combination we use. In addition to popular P2P IR testbeds comprising ≈1.6 million documents listed from [11], we also evaluate our approach on a meta-search task on the FedWeb 2013 dataset [6] (13k documents, 200 queries) consisting of results across 157 search engines, each of which are modeled as a separate peer. Since the result trends across methods were found to be consistent across varying values of pre-specified percentages of peers to be selected at each super-peer, we will report the average of results from 10 settings with the selectivity parameter varying from 5% to 50% in increments of 5%.

6. EXPERIMENTAL RESULTS

P2P IR Testbeds: Table 3 summarizes the results on the IR testbeds viz., DLWOR, DLWR, ASISWOR and ASISWR (details in [11]); DL* and ASIS* simulate digital library and file-sharing scenarios respectively. We report precision, recall and MAP figures evaluated over the top-1000 documents, for each technique-testbed combination. While top-1000 is usually realistic for getting a subset of relevant documents in P2P IR, fewer documents are usually expected to be utilized in scenarios involving manual perusal of results; we thus report the Precision at top-10 results too, in the table. The top and bottom parts of the table illustrate the document retrieval and standard resource selection methods respectively; each value that is seen to be better than the best value for the [metric,testbed] combination in the other category is boldfaced. Additionally, the best value across categories is underlined as well. We performed bootstrap 2-paired t-tests to analyze whether the performance improvements of document retrieval methods were statistically significant; results of such tests are also indicated in the table. It may be seen from the table that the document retrieval methods perform better in every testbed and evaluation metric, barring one case where Taily is seen to outperform others. The last row counts the number of document retrieval methods (out of the 7 being studied) that perform either equally or better than the best performing general resource selection method under study, on the corresponding evaluation metric. As may be seen, on an average, roughly 40% of document retrieval methods are seen to outperform the best performing resource selection methods, across metrics. This establishes that document retrieval methods are very effective for resource selection in the clustered P2P IR architecture, and should be preferred to classical resource selection methods designed for general P2P IR.

Federated Web Search Testbed: Table 4 summarizes

[2] http://trec.nist.gov/data/webmain.html

Table 3: Results on P2P IR Testbeds (○ & • indicate statistical significance at $p < 0.05$ and $p < 0.01$ respectively)

Method	DLWOR				DLWR				ASISWOR				ASISWR			
	Prec	Recall	MAP	Pr@10	Prec	Recall	MAP	Pr@10	Prec	Recall	MAP	Pr@10	Prec	Recall	MAP	Pr@10
TF-IDF	**0.0293**$^{\bullet}$	**0.5244**	**0.0911**	**0.1792**	**0.0256**	**0.4740**	**0.0307**$^{\bullet}$	0.0253	0.0189	0.3391	0.0531	0.1342	0.0191	0.3741	0.0194	0.0150
BM25	0.0177	0.2878	0.0559	0.1108	0.0170	0.3019	0.0241	0.0267	0.0253	0.4516	0.0719	0.1602	**0.0199**	**0.3919**	0.0195	0.0143
LM	0.0281	0.4889	**0.0896**	**0.1805**	**0.0259**	0.4591	**0.0349**$^{\bullet}$	**0.0357**$^{\circ}$	0.0256	0.4427	**0.0742**	**0.1727**$^{\circ}$	**0.0217**$^{\circ}$	**0.4009**$^{\circ}$	**0.0232**	**0.0157**
BB2	**0.0290**$^{\circ}$	**0.5238**	**0.0925**	**0.1765**	0.0250	0.4721	**0.0300**$^{\circ}$	0.0242	**0.0263**	**0.4689**	**0.0757**$^{\circ}$	**0.1703**$^{\circ}$	0.0188	0.3713	0.0194	0.0150
Inexp_B2	**0.0286**	0.5150	0.0884	0.1740	0.0247	0.4625	**0.0289**	0.0242	0.0257	0.4563	**0.0742**	**0.1703**$^{\circ}$	0.0190	0.3693	0.0198	0.0152
Inexp_C2	**0.0286**	0.5136	**0.0895**	0.1740	0.0247	0.4638	**0.0288**	0.0245	0.0256	0.4537	**0.0741**	**0.1710**$^{\circ}$	0.0191	0.3686	0.0199	0.0153
InL2	**0.0291**$^{\bullet}$	0.5130	0.0887	**0.1775**	**0.0255**	0.4658	**0.0307**$^{\bullet}$	0.0257	**0.0260**	0.4628	**0.0741**	**0.1680**	0.0192	0.3771	0.0196	0.0150
CVV	0.0266	0.4961	0.0842	**0.1757**	0.0216	0.4254	0.0236	0.0198	0.0252	0.4594	0.0714	0.1635	0.0191	**0.3846**	0.0188	0.0147
CORI	0.0280	0.4983	0.0845	0.1697	0.0243	0.4532	**0.0282**	0.0242	**0.0260**	**0.4645**	**0.0737**	**0.1640**	0.0186	0.3694	0.0188	0.0143
vGlOSS	0.0245	0.4510	0.0777	0.1643	0.0212	0.3970	0.0275	**0.0288**	0.0226	0.4089	0.0705	0.1593	**0.0197**	0.3820	**0.0215**	0.0152
KL	0.0172	0.2727	0.0453	0.1323	0.0155	0.2461	0.0175	0.0275	0.0188	0.3368	0.0530	0.1347	0.0170	0.3221	0.0175	0.0155
Taily	**0.0282**	**0.5216**	**0.0894**	0.1605	**0.0251**	**0.4777**	0.0277	0.0238	0.0258	0.4606	0.0705	0.1583	0.0193	0.3783	0.0204	**0.0173**
	5	**2**	**4**	**4**	**3**	**-**	**6**	**1**	**2**	**1**	**5**	**5**	**2**	**2**	**1**	**-**

Table 4: Results on FedWeb2013 Testbed

Method	Prec	Recall	MAP	Pr@10	nDCG@10
TF-IDF	**0.0626**$^{\bullet}$	**0.5858**$^{\bullet}$	**0.2526**$^{\bullet}$	**0.4285**$^{\bullet}$	**0.2721**$^{\bullet}$
BM25	0.0178	0.2223	0.1161	0.1842	0.1159
LM	0.0475	0.4260	0.1552	0.3036	0.1951
BB2	**0.0641**$^{\bullet}$	**0.6047**$^{\bullet}$	**0.2610**$^{\bullet}$	**0.4430**$^{\bullet}$	**0.2886**$^{\bullet}$
Inexp_B2	**0.0622**$^{\bullet}$	**0.5835**$^{\bullet}$	**0.2490**$^{\bullet}$	**0.4117**$^{\circ}$	**0.2657**$^{\bullet}$
Inexp_C2	**0.0622**$^{\bullet}$	**0.5835**$^{\bullet}$	**0.2490**$^{\bullet}$	**0.4113**$^{\circ}$	**0.2655**$^{\bullet}$
InL2	**0.0615**$^{\bullet}$	**0.5692**$^{\bullet}$	**0.2458**$^{\bullet}$	**0.4203**$^{\bullet}$	**0.2680**$^{\bullet}$
CVV	0.0496	0.4627	0.1566	0.3184	0.2010
CORI	0.0431	0.3953	0.1389	0.2922	0.1656
vGlOSS	0.0340	0.3155	0.1050	0.2381	0.1454
KL	0.0234	0.2071	0.0508	0.1599	0.1225
Taily	**0.0580**	**0.5419**	**0.2228**	**0.3808**	**0.2198**
	5	**5**	**5**	**5**	**5**

the corresponding results on the federated web search testbed, in the same format as for P2P IR testbeds. The results on this testbed also confirm the superior performance of the document retrieval methods for resource selection in the clustered P2P IR architecture in the meta-search environments. It is notable that five of seven document retrieval methods outperform the best performing general resource selection method, on each metric.

7. CONCLUSIONS AND FUTURE WORK

We considered the problem of resource selection in the clustered P2P IR architecture. We postulated an enhanced applicability of document retrieval methods for resource selection in the target architecture, given the content homogeneity among intra-peer clusters at the resource selection layer. We empirically benchmarked well-known document retrieval methods against the state-of-the-art resource selection methods designed for general P2P IR. An extensive analysis of retrieval effectiveness over P2P IR and federated-search testbeds using classical IR evaluation metrics validate our hypothesis convincingly, with document retrieval methods consistently outperforming others. This establishes that document retrieval methods ought to be the preferred choice for resource selection in clustered P2P IR environments. Our empirical analysis indicates that the language modeling approach (LM) should be the preferred resource selection method for both P2P IR and Federated Web Search scenarios; further, BB2 is seen to be competitive in P2P IR.

From a future work perspective, our results indicate a potential convergence of the document retrieval and resource selection tasks in the clustered P2P IR architecture so that future advancements in document retrieval may be effectively leveraged to achieve corresponding gains in resource selection on the target framework. With the improved accuracy of document retrieval techniques being apparent from our study, their indexing overhead and messaging costs need to be analyzed for fast uptake.

8. REFERENCES

[1] R. S. Alkhawaldeh and J. M. Jose. Experimental study on semi-structured peer-to-peer information retrieval network. In *In CLEF 2015, Toulouse, France, September 8-11, 2015, Proceedings*, pages 3–14. 2015.

[2] R. Aly, D. Hiemstra, and T. Demeester. Taily: Shard selection using the tail of score distributions. In *SIGIR*, 2013.

[3] G. Amati and C. J. Van Rijsbergen. Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM TOIS*, 20(4):357–389, 2002.

[4] R. Baeza-Yates, B. Ribeiro-Neto, et al. *Modern information retrieval*, volume 463. ACM press New York, 1999.

[5] J. P. Callan, Z. Lu, and W. B. Croft. Searching distributed collections with inference networks. In *SIGIR*, 1995.

[6] T. Demeester, D. Trieschnigg, D. Nguyen, and D. Hiemstra. Overview of the trec 2013 federated web search track. In *Proceedings of the TREC*, pages 1–11, 2013.

[7] N. Fuhr. A decision-theoretic approach to database selection in networked ir. *ACM TOIS*, 17(3):229–249, 1999.

[8] L. Gravano, H. García-Molina, and A. Tomasic. Gloss: Text-source discovery over the internet. *ACM Trans. Database Syst.*, 24(2):229–264, June 1999.

[9] D. Hiemstra. *Using language models for information retrieval*. Taaluitgeverij Neslia Paniculata, 2001.

[10] I. A. Klampanos and J. M. Jose. An evaluation of a cluster-based architecture for peer-to-peer information retrieval. In *DEXA*, pages 380–391, 2007.

[11] I. A. Klampanos, V. Poznański, J. M. Jose, P. Dickman, and E. H. Road. A suite of testbeds for the realistic evaluation of peer-to-peer information retrieval systems. In *ECIR*, 2005.

[12] J. H. Lee. Analyses of multiple evidence combination. *SIGIR Forum*, 31(SI):267–276, July 1997.

[13] J. Lu and J. Callan. Federated search of text-based digital libraries in hierarchical peer-to-peer networks. In *Proceedings of ECIR'05*, pages 52–66, Berlin, Heidelberg, 2005. Springer-Verlag.

[14] M. Melucci and A. Poggiani. A study of a weighting scheme for information retrieval in hierarchical peer-to-peer networks. In *Proceedings of ECIR*, pages 136–147, 2007.

[15] S. E. Robertson, S. Walker, S. Jones, M. M. Hancock-Beaulieu, M. Gatford, et al. Okapi at trec-3. *NIST*, 109:109, 1995.

[16] L. Si and J. Callan. Unified utility maximization framework for resource selection. In *CIKM*, pages 32–41. ACM, 2004.

[17] L. Si, R. Jin, J. Callan, and P. Ogilvie. A language modeling framework for resource selection and results merging. In *CIKM*, pages 391–397. ACM, 2002.

[18] A. S. Tigelaar, D. Hiemstra, and D. Trieschnigg. Peer-to-peer information retrieval: An overview. *ACM TOIS*, 2012.

[19] J. Xu and W. B. Croft. Cluster-based language models for distributed retrieval. In *SIGIR*, 1999.

[20] B. Yuwono and D. L. Lee. Server ranking for distributed text retrieval systems on the internet. In *DASFAA*, 1997.