Clustering-based Query Routing in Cooperative Semi-structured Peer to Peer Networks

Rami S. Alkhawaldeh[†]
School of Computing Science
University of Glasgow
University Avenue, G12 8QQ
Glasgow, United Kingdom
r.alkhawaldeh.1@research.gla.ac.uk

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

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

[†]Department of Computer Information Systems, The University of Jordan Aqaba, 77110, Jordan

***Abstract*—We consider the problem of resource selection in clustered Peer-to-Peer Information Retrieval (P2P IR) networks with cooperative peers. The clustered P2P IR framework presents a significant departure from general P2P IR architectures by employing clustering to ensure content coherence between resources at the resource selection layer, without disturbing document allocation. We propose that such a property could be leveraged in resource selection by adapting well-studied and popular inverted lists for centralized document retrieval. Accordingly, we propose the Inverted PeerCluster Index (IPI), an approach that adapts the inverted lists, in a straightforward manner, for resource selection in clustered P2P IR. IPI also encompasses a strikingly simple peer-specific scoring mechanism that exploits the said index for resource selection. Through an extensive empirical analysis on P2P IR testbeds, we establish that IPI competes well with the sophisticated state-of-the-art methods in virtually every parameter of interest for the resource selection task, in the context of clustered P2P IR.**

***Keywords*—Clustering Peers; Semi-structured; Peer to Peer; Information Retrieval; Query Routing; Evaluation.**

I. Introduction

Resource Selection in P2P IR systems is the problem of selecting a subset of relevant peers that are most promising with respect to a user query [6], [7]. Co-operative P2P IR systems are those where the broker that is responsible for resource selection is free to acquire information from peers it manages, and uses such information to prioritize resources for a query. The results from selected peers are then merged to create a final result list. Resource selection is a critical component in P2P IR; excluding relevant peers in the resource selection stage would inevitably lead to less accurate IR results.

This paper focuses on the problem of resource selection in co-operative clustered P2P IR networks. Co-operative clustered P2P IR networks use clustering to build semantically coherent peer-clusters [8]. As we will see, the clustered P2P IR architecture involves a two-level clustering so that coherent documents are first grouped within each peer, followed by aggregating coherent clusters from across peers. The resource selection task is then applied to such homogeneous cluster groups involving multiple peers. This is in sharp contrast to a general P2P IR framework where the resource selection layer needs to work across diverse resources. For example, in a typical case for federated search over various news agencies, each news agency managed by a separate peer would comprise documents as diverse as the entire corpus. This property of general P2P IR has led to development of methods that model and exploit distributional information (e.g., variance) of terms across peers, in their scoring process. Much like in traditional IR, P2P IR resource selection works by computing a query-specific score for each peer, followed by choosing the top-scoring peers to route the query to. Reliance on inter-peer distributional information in computing peer-specific scores induces a dependency at the resource selection layer; for example, updates within a document collection in one peer would require a re-computation of the distributional information held at the resource selection level, and would result in changes in the score of *other* peers, for a subsequent query[1]. This restricts parallelism between the update processing and scoring processes at the resource selection layer. We postulate that the content homogeneity at the resource selection layer in clustered P2P IR makes such sophistication an overkill, and that the semantic coherence at the resource selection layer could enable us to fall back on much simpler peer-specific models for resource selection. Our main contributions are as follows:

- We posit that simplistic word frequency based models would be able to leverage the content homogeneity clustered P2P IR frameworks to accurately perform resource selection. Accordingly, we adopt conventional inverted indexes from IR literature for resource selection in clustered P2P IR and propose IPI, a remarkably simple resource selection method for clustered P2P IR.
- Through an extensive empirical evaluation on classical P2P IR testbeds, we establish that IPI competes with sophisticated resource selection methods for virtually every parameter of interest.

This paper is organised as follows: We first outline related

[1]The analogous dependency in the clustered architecture is the change in clustering assignments, which are handled at the framework level beneath the resource selection layer.

TABLE I: Related Work Overview (α, β are method-specific parameters)

Method	Info Per Peer (IPP)	IPP Size	Resource Score Computation
CVV[1]	$\forall w, DF_P(w)$	$\mathcal{O}(W)$	$\propto \sum_{q \in Q} CVV(q) \times DF_P(q)$ where $CVV(w)$ is the variance of the distribution of w across peers
Taily[2]	For every word, Expected Frequency & Expected Variance	$\mathcal{O}(W)$	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
CORI [3]	$\forall w, DF_P(w)$	$\mathcal{O}(W)$	$\propto \sum_{q \in Q} \frac{DF_P(q)}{DF_P(q)+\alpha+\beta \times \#words\_in\_P}$
KL[4]	$\forall t \in P, \forall w, Prob(w\|t)$	$\mathcal{O}(WT)$	Multiple topic-specific langauge models in a peer; query is matched against LMs using KL-divergence and assigned to the collections with best matching topics.
vGlOSS[5]	$\forall w, DF_P(w) \& \#Docs_P(w)$	$\mathcal{O}(W)$	Estimate of the number of documents that the peer would return assuming terms in the query co-occur in documents
IPI (Our)	$\forall w, IPI[w].P$	$\mathcal{O}(W)$	$\sum_{q \in Q} IPI[q].P$

work in the P2P IR resource selection and our target P2P IR architecture. We then describe our resource selection method, followed by an empirical analysis and conclusions.

II. Related Work

In cooperative P2P IR networks, brokers maintain peer-level collection statistics, that summarize the content at each peer, to determine the relevant peers for the given query [9]. Resource selection methods differ in the kind of peer-level statistics maintained, and the scoring methodology used to assess the estimated relevance of peers to a query. As discussed earlier, state-of-the-art methods maintain and exploit cross-peer term distribution information to score peers.

A terse summary of well-known resource selection methods in cooperative environments used in this study appears in Table I. We use three notations, (1) $DF_P(w)$ that denotes the sum of document frequencies (or tf.idf) of w across documents in the peer P, (2) $\#Docs_P(w)$ that simply counts the number of documents that contain w, and (3) $p(w|.)$ the probability of w across documents for each topic in P. Table I illustrates that all techniques use some form of cross-peer distributional information; examples include word-specific peer frequencies in CORI and variance in CVV and Taily.

III. Clustered P2P IR Architecture

We now describe the well-studied clustered P2P IR architecture (used in [8], [10] and various others), our target architecture. This leverages clustering methods so that the resource selection can be done at the level of coherent groups of documents from across peers. It uses two levels of clustering:

- **Stage 1, Intra-peer Clustering:** Each peer clusters the set of documents it manages using a text clustering method such as K-Means so that they are grouped into multiple semantically coherent clusters; we call these as *peer-clusters*. The peer-clusters represent different topics within a peer (e.g., sports). The topic centroids are used to do further processing in Stage 2.
- **Stage 2, Super-Peers:** The topical centroids of peer-clusters from across peers are further clustered into a number of clusters. Hence, centroids corresponding to similar topics from separate peers are expected to be aggregated into a robust topical coherent representation for resource selection at super-peer level. At runtime, resource selection is performed at the level of each such cluster of peer-clusters. Each of the clusters output from this operation would be managed by a *super peer* that is responsible for resource selection.

The 2-stage clustering process ensures that routing decisions can be made at the level of super-peers that manage coherent content internally, while not disturbing the document assignment to peers; this likens the scenario to a domain-specific search engine at each super-peer.

Each super-peer would use the resource selection algorithm chosen by the designer to route the query to a subset of the peers among those whose peer-clusters it manages; there is no inter-super-peer communication, unlike the unclustered ultra-peer architecture [11]. As outlined, each super-peer manages information about multiple clusters from across peers. We will denote the k^{th} cluster from peer P_i as P_i^k. For every super-peer S_j, C_{S_j} denotes the set of clusters that are managed by S_j. Due to the clustering-based construction, C_{S_j} could contain zero, one or multiple clusters from a specific peer. We use a centroid-based representation throughout; thus, $Cd(P_i^k)$ denotes the centroid of the documents within the k^{th} cluster in the i^{th} peer. Specifically, the centroid in the vector space has as many entries as there are words in the vocabulary, the entry for each word takes the average of the value of the word across the components.

$$Peer\ Centroids : Cd(P_i^k)[w] = \frac{\sum_{d \in P_i^k} tf.idf(w,d)}{\#docs\ in\ P_i^k}$$

$$Super{-}peer\ Centroids : Cd(C_{S_j})[w] = \frac{\sum_{P_x^y \in C_{S_j}} Cd(P_x^z)[w]}{\#clusters\ in\ C_{S_j}}$$

where $tf.idf(w, d)$ denotes the tf-idf score of the word w in document d.

IV. Inverted PeerCluster Index

Motivation: Our approach seeks to exploit the content coherence at the peer-cluster level in the clustered architecture to devise a simple scoring method that uses the conventional inverted indexing approach for Information Retrieval. In particular, we expect that the clustering step mitigates the heterogeneity of content at the super-peer level to an extent so that simple word-frequency scores are meaningful and informative enough. Content diversity in unclustered environments can be thought of as contaminating simple frequentist statistics (due to mixing of diverse semantics); this warrants the usage of cross-peer information (e.g., estimates of variance) as is commonly employed in state-of-the-art resource selection methods surveyed in Section II. On the other hand, the clustering inherent in the clustered P2P architecture is itself a model of cross-peer information; thus, we consider exploiting that property of the architecture to bypass modeling cross-peer distributional information again at the resource selection layer. We will now outline the construction of our inverted index based resource selection method.

Usage: It may be noted that the two-layered clustered architecture allows for resource selection of two levels: one where a subset of super-peers may be chosen using the super-peer centroids, and another where a subset of peer-clusters may be chosen at each super-peer based on the peer-cluster centroids. We will focus on the latter, assuming that the query is made available to all super-peers with results across super-peers being merged centrally; this also allows for a fair comparison against single-level P2P networks.

Inverted PeerCluster Index: The inverted peercluster index at any super-peer is simply an inverted index (i.e., word-level lists) over the peers; each peer is tagged with a score that is aggregated across the peer-clusters from the peer. The word-level index for the word w at a super-peer S_j would contain 2-tuples in the form of $[peer, score]$ entries:

$$IPI(S_j)[w] = \{\Big[P_x, \sum_{P_x^y \in C_{S_j}} Cd(P_x^y)[w]\Big] | P_x : \exists y, P_x^y \in C_{S_j}\}$$

Thus, if S_2 manages two peer-clusters from P_3, the entry for word w for P_3 would be the sum of the entries for the word in the two $Cd(P_3^i)$s within C_{S_2}. Despite the lists entries being peer-specific, we call it a peer-cluster index since the corresponding scores are computed by aggregating across only those peer-clusters that belong to the super-peer (and not across *all* documents in the peer). We will denote IPI much like an associative array where $L.P_x$ denotes the score for P_x in the list L. Given a query Q containing terms $\{q_1, q_2, \ldots, q_l\}$, we then score peers within S_j as follows.

$$Score(P_x, Q) = \begin{cases} 0, \text{if } \exists q_i,\ IPI(S_j)[q_i].P_x = \phi \\ \sum_{q_i \in Q} IPI(S_j)[w].P_x, \text{otherwise} \end{cases}$$

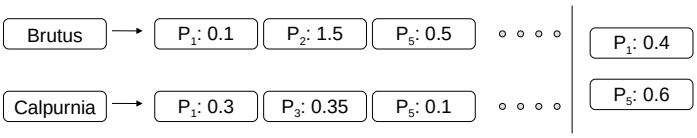


Fig. 1: IPI Example; Scoring for Query *'Brutus Calpurnia'* on the right

Thus, only those peers who have an entry in the list corresponding to each query term are *eligible*; the eligible peers are then scored using a sum-based aggregation of corresponding entries. The eligibility condition is a scoring adaptation targeted specifically at the P2P IR use case. It may be noted that the messaging cost is a step function, with a fixed cost (of communicating the query, and collecting the results) incurred for every selected resource. A discrete eligibility function serves to exclude potentially irrelevant resources upfront towards reducing messaging costs.

Depending on the budget constraint (in terms of computational expense), the peer-clusters with the top scores are chosen for S_j to route the query to. Declaratively, if k peer-clusters are to be chosen,

$$Top\text{-}k@(S_j, Q) = \underset{R \subseteq C_{S_j}, |R|=k}{\arg\max} \sum_{P_x \in R} Score(P_x, Q)$$

The typical resource allocation algorithm chooses k as a specified percentage of peers according to the selection mechanism adopted by that approach. If there are fewer eligible peers than the specified percentage, only the eligible ones are selected. The fraction operates as a meta-parameter to the selection algorithm. The selected peers, then process the query in a cluster-agnostic manner, following which their results are merged.

Example: Figure 1 shows an example structure of the IPI index (sorted by peer id) in a super-peer, and the aggregated scores for a query *Brutus Calpurnia* to illustrate the working of the IPI resource selection approach. P_2 and P_3 are ineligible since they occur only in one of the lists; P_4, on the other hand, does not occur in either list.

Setting up: The simplicity of the IPI formulation is also reflected in the ease of setting up the index. A step involving frequency counting of words is the only overhead in the set-up phase, in contrast with the techniques in Table I.

V. Experimental Evaluation

Experimental Setup: We use three P2PIR testbed categories, DL*, ASIS*, and U*, with 1500, 11680 and 11680 peers respectively, which are derived from the WT10g collection [12]; the WR and WOR variants indicate where there is content replication across peers or not. The standard query set of TREC topics 451-550[2] is used and we report performance measurements averaged across the 100 queries in the set. For

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

TABLE II: IPI Retrieval effectiveness at 10% of Selected Peers

DL*	**DLWOR Testbed**				**DLWR Testbed**			
Method	Precision	Recall	P@10	MAP	Precision	Recall	P@10	MAP
Flooding	0.02866	0.54790	0.16900	0.08659	0.02089	0.42564	0.01837	0.02232
IPI	**0.02461**	**0.47601**	**0.188**	**0.09358**	**0.0229**	**0.4534**	**0.023**	**0.02662**
CVV	0.02435	0.44472	0.182	0.08281	0.02158	0.41428	0.02	0.0239
Taily	**0.02769**	**0.50463**	0.182	**0.09493**	**0.02523**	**0.46911**	**0.026**	**0.02978**
CORI	**0.02686**	0.45697	0.177	0.0864	**0.0256**	0.44471	**0.027**	**0.03171**
KL	0.01105	0.1513	0.109	0.02782	0.0104	0.13508	**0.034**	0.01295
vGIOSS	0.02208	0.38807	0.171	0.07616	0.0183	0.33898	**0.034**	**0.02915**
RW	0.01455	0.18186	0.134	0.03856	0.01516	0.13561	**0.042**	0.0122
IPI Rank	3	2	1	2	3	2	6	4
ASIS*	**ASISWOR Testbed**				**ASISWR Testbed**			
Method	Precision	Recall	P@10	MAP	Precision	Recall	P@10	MAP
Flooding	0.02532	0.46296	0.16400	0.07122	0.01635	0.33847	0.01414	0.01732
IPI	**0.02469**	**0.45859**	**0.163**	**0.07047**	**0.02117**	**0.41612**	**0.016**	**0.01998**
CVV	**0.02508**	0.45601	**0.166**	**0.07227**	0.02053	0.41176	**0.016**	0.01954
Taily	**0.02613**	**0.45867**	0.159	**0.07103**	0.02088	0.39249	**0.02**	**0.02218**
CORI	**0.02636**	**0.46426**	0.161	**0.0743**	0.02066	0.39906	0.014	**0.02014**
KL	0.01525	0.25003	0.121	0.04275	0.01565	0.27106	0.015	0.01721
vGIOSS	0.01966	0.35336	0.153	0.06321	0.01901	0.35926	**0.016**	**0.02386**
RW	0.0131	0.16287	0.111	0.03037	0.01451	0.17353	**0.019**	0.01084
IPI Rank	4	3	2	4	1	1	3	4
U*	**UWOR Testbed**				**UWR Testbed**			
Flooding	0.02764	0.49910	0.21200	0.10581	0.02269	0.42657	0.01400	0.02331
IPI	**0.02582**	**0.46564**	**0.211**	**0.10022**	**0.02269**	**0.42657**	**0.014**	**0.02331**
CVV	0.02158	0.38	0.186	0.0862	0.01932	0.36159	0.012	0.01926
Taily	**0.02781**	**0.48616**	0.183	**0.10931**	**0.02563**	**0.45538**	**0.014**	**0.02751**
CORI	**0.0273**	**0.46807**	0.191	0.09365	**0.02674**	**0.46429**	**0.014**	**0.02969**
KL	0.01295	0.20201	0.122	0.0478	0.01319	0.24118	**0.015**	0.01491
vGIOSS	0.01859	0.34657	0.157	0.06733	0.01906	0.35583	**0.016**	**0.02367**
RW	0.01405	0.16086	0.117	0.03062	0.01463	0.17359	**0.02**	0.00832
IPI Rank	3	3	1	2	3	3	4	4

all methods in the evaluation, we use the COMBMNZ merging algorithm [13] for merging the results of selected resources, which are then evaluated on standard IR evaluation metrics. For baselines, we used the resource selection algorithms from Table I.

A. Experimental Results and Discussion

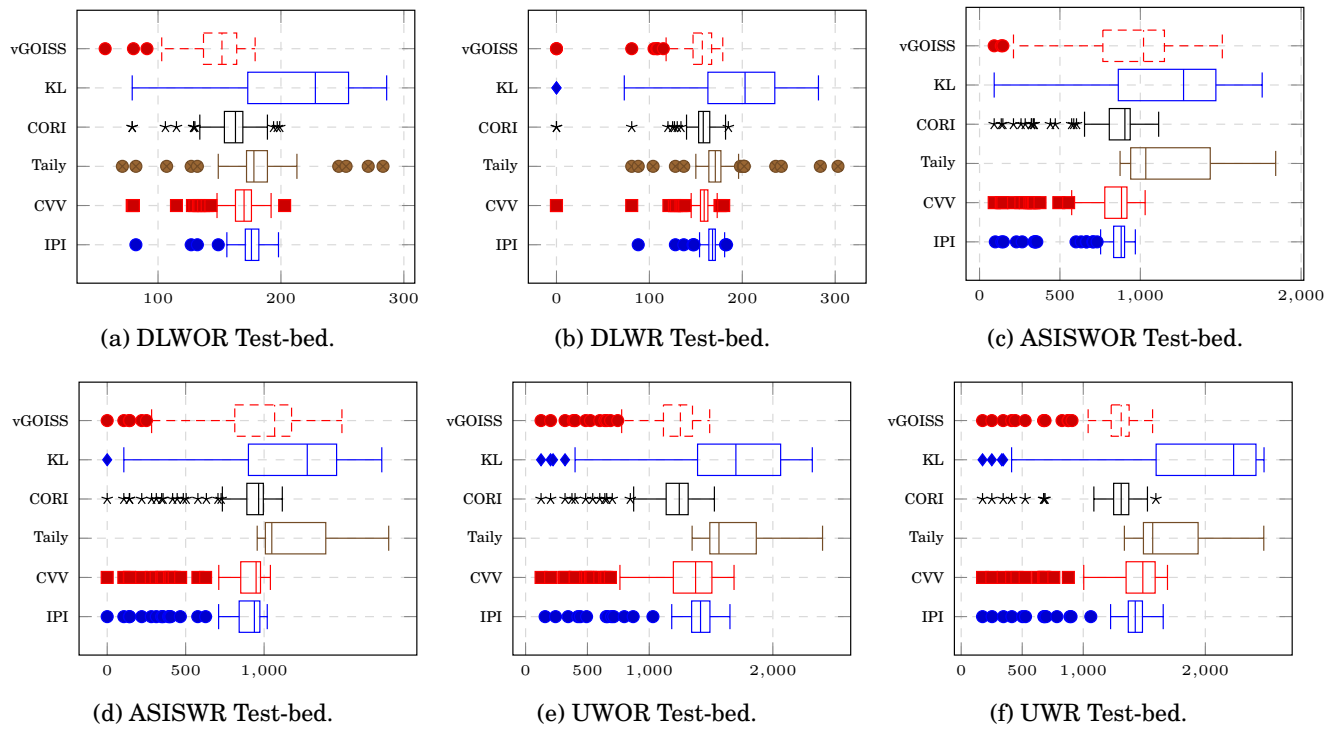
Two core dimensions are of interest; effectiveness as measured by the quality of results on IR evaluation metrics, and the efficiency as measured by the messaging costs indicated by the number of peers chosen. We analyze them separately.

Effectiveness/Result Quality: Table II lists Precision, Recall, P@{10, 30, 100} and MAP measures evaluated at the top-1000 in the merged list. As may be seen, IPI performs on par with state-of-the-art resource selection methods across the various testbeds. The rank of IPI listed under each scenario-metric combination summarizes the competitiveness of IPI. This outlines that IPI is very competitive with state-of-the-art resource selection methods in the clustered P2P IR setting.

Efficiency/Messaging: Another core parameter of interest in the P2P IR resource selection is the messaging cost; Figure 2 shows illustrating messaging cost distributions. To summarize the charts, IPI is seen to incur mostly similar messaging costs to CORI, Taily, and CVV; these four methods were seen to be those that scored high on accuracy measures in our effectiveness study. KL is seen to incur slightly higher messaging costs on an average, while vGlOSS is seen to be slightly more efficient in messaging over IPI, CORI, Taily, and CVV.

VI. Conclusions

In this paper, we addressed the problem of resource selection in the clustered P2P IR architecture. We observed that clustering within the architecture may be seen as a model of cross-peer information and discern that sophisticated



(a) DLWOR Test-bed. (b) DLWR Test-bed. (c) ASISWOR Test-bed.

(d) ASISWR Test-bed. (e) UWOR Test-bed. (f) UWR Test-bed.

Fig. 2: Efficiency Analysis: Messaging Costs

models that maintain and exploit cross-peer term occurrence distributional metrics at the resource selection layer would not be necessary for this architecture. Accordingly, we outlined a simple index-based method for resource selection, called IPI, adapting inverted indexes from centralized IR literature. Through an extensive analysis on P2P IR testbeds, our method, IPI, is seen to be as good as state-of-the-art resource selection methods designed for general P2P IR architecture, both in terms of accuracy as well as messaging costs; this establishes IPI as a simple and effective resource selection method for clustered P2P IR. The independence between the scores of peers due to non-usage of term occurrence distributional information is an attractive feature in IPI, making it a natural choice for resource selection in clustered P2P IR.

In future work, we plan to explore methods to adapt the IPI approach to un-co-operative environments by devising sampling strategies that could identify intra-peer clusters, so that the notion of peer clusters may extend naturally.

References

[1] B. Yuwono and D. L. Lee, "Server ranking for distributed text retrieval systems on the internet," in *DASFAA*, 1997, pp. 41–50.

[2] R. Aly, D. Hiemstra, and T. Demeester, "Taily: shard selection using the tail of score distributions," in *SIGIR*. ACM, 2013, pp. 673–682.

[3] J. P. Callan, Z. Lu, and W. B. Croft, "Searching distributed collections with inference networks," in *Proc. of SIGIR*, 1995, pp. 21–28.

[4] J. Xu and W. B. Croft, "Cluster-based language models for distributed retrieval," in *Proc. of SIGIR*. ACM, 1999, pp. 254–261.

[5] L. Gravano, H. García-Molina, and A. Tomasic, "Gloss: Text-source discovery over the internet," *ACM Trans. Database Syst.*, vol. 24, no. 2, pp. 229–264, Jun. 1999.

[6] A. S. Tigelaar, D. Hiemstra, and D. Trieschnigg, "Peer-to-peer information retrieval: An overview," *ACM Trans. Inf. Syst.*, vol. 30, no. 2, pp. 9:1–9:34, May 2012.

[7] J. Lu and J. Callan, "Federated search of text-based digital libraries in hierarchical peer-to-peer networks," in *ECIR*, 2005, pp. 52–66.

[8] I. A. Klampanos and J. M. Jose, "An architecture for peer-to-peer information retrieval," in *Proc. of SIGIR*. New York, NY, USA: ACM, 2003, pp. 401–402.

[9] S. Richardson and I. J. Cox, "Estimating global statistics for unstructured p2p search in the presence of adversarial peers," in *Proc. of SIGIR*, 2014, pp. 203–212.

[10] 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*, 2015, pp. 3–14.

[11] E. Di Buccio, I. Masiero, and M. Melucci, "Improving information retrieval effectiveness in peer-to-peer networks through query piggybacking," in *Research and Advanced Technology for Digital Libraries*. Springer, 2009, pp. 420–424.

[12] 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 *Proc. of ECIR*, 2005, pp. 38–51.

[13] J. H. Lee, "Analyses of multiple evidence combination," *SIGIR Forum*, vol. 31, no. SI, pp. 267–276, Jul. 1997.