On Multiword Entity Ranking in Peer-to-Peer Search

Yuval Merhav
Information Retrieval Laboratory
Computer Science Department
Illinois Institute of Technology
yuval@ir.iit.edu

Ophir Frieder
Georgetown University
and
IIT Information Retrieval Laboratory
ophir@cs.georgetown.edu

ABSTRACT
Previously [2], we postulated the advantage of using entity extraction to implement a new Peer-to-Peer (P2P) search framework for reducing network traffic and providing a trade off between precision and recall. We now propose an entity ranking method designed for the 'short documents' characteristic of P2P, which significantly improves both precision and recall in 'top results' P2P search. We construct a dynamic entity corpus using n-grams statistics and metadata, study its reliability, and use it to identify correlations between user query terms.

Categories and Subject Descriptors:
H.3.3 [**Information Storage and Retrieval**]: Information search and retrieval

General Terms: Performance, Experimentation.

1. INTRODUCTION
Peer-to-peer (P2P) file sharing systems are distributed networks in which peers exchange information directly with each other. Every peer has a set of shared files, and each file is described by a list of terms called a descriptor, which we view as a short document. In most P2P file-sharing systems, to find a file, users issue a query, terms in the query are independently matched against descriptors of other peers, and only those that match all query terms are returned to the user (conjunctive queries (CQ)). The main disadvantage is clear: no attention is given for the semantics that many multiple terms constitute of.

We identify multiword named entities (of any type) by computing the statistical correlation between terms in each n-gram accumulated over P2P data. N-grams that include terms with high correlation between them generate a named entity corpus, and are used to parse each query to its correct entities by a simple matching of the n-grams appearing in the query with the corpus entities. To evaluate our method, we compare a proposed statistical function against a well studied one, and evaluate both over four well known performance measures. We report a minor difference between the two, but a significant improvement over standard conjunctive queries ranking. We then enrich the entity corpus using metadata and show an additional improvement.

2. THE STATISTICAL MODEL
Each file for each peer contains a descriptor which describes the file (e.g., Red Hot Chili Peppers, Greatest Hits). We construct an n-gram table that includes unigrams, bigrams, trigrams, and 4grams, based on the large collection of descriptors that peers hold. Each n-gram is stored with its frequency.

SIGIR'08, July 20-24, 2008, Singapore.
ACM 978-1-60558-164-4/08/07.

After constructing the n-gram table, along with their frequencies, we apply a statistical measure method called '*the fair Symmetric Conditional Probability' (SCP)* that tests the statistical correlation between n terms [1].

$$SCP((w_1...w_n)) = \frac{p(w_1...w_n)^2}{AVP}$$

$$AVP = \frac{1}{n-1}\sum_{i=1}^{i=n-1} p(w_1...w_i)\cdot p(w_{i+1}...w_n)$$

where n is the number of terms $(w_1...w_n)$, and $p(w_1...w_n)$ is the probability of terms $(w_1...w_n)$ to appear together in the given order. AVP is computed by averaging over all possible partitions.

For comparison, we use an additional correlation function called *'Exponential Frequency Growth'* (EFG): [2]

$$EFG(w_1,...,w_n) = \frac{f(w_1...w_n)^n}{AVG(f(w_1),...,f(w_n))}$$

where $f(w_i)$ is the frequency of term w_i in the collection. $AVG(f(w_1),...,f(w_n))$ is the average frequency of the terms $(w_1,...,w_n)$ (each term frequency is independent).

3. METADATA ENRICHMENT
Our experiments showed that SCP and EFG are both successful statistical functions to identify multi word entities. However, due to the inconsistent distribution of different entities, both functions fail to identify exceptions (e.g., rare entities). Thus, we added an optional enhanced technique that exploits the presence of metadata in many shared files (e.g. artist name, album name, etc.), and use it to enrich the entity corpus.

4. CORPUS RELIABILITY
4.1 SCP/EFG Efficiency
We tested both SCP and EFG performance on 1,000,000 file descriptors crawled from LimeWire's Gnutella system, using IR-Wire, a publicly available research tool [3]. We also used a database of 20,687 song artists (performers and songwriters) and 55,794 songs (the data were collected by www.secondhandsongs.com), likewise publicly available.

After computing n-grams and frequencies accumulated from the collection, we separately applied the SCP and EFG functions to every n-gram in the table. Ideally, real entities would get higher values on average than non-entities; therefore, we compared the average of our known entities SCP/EFG values, with the average of the entire n-grams SCP/EFG values. Clearly, the n-gram collection has many true multiword entities that are not present in our entity

database, and therefore, we expect those entities to have high values as well. However, we assume that most of the n-grams are not entities, and we expect the average of the entire collection to be significantly lower than the average of a small group of known multiword entities. Table 1 confirms our assumption; we note a significant difference between the statistical functions average values for the known entities and the entire n-gram collection. For example, the EFG average value of 2,500 bigrams that were identified as 2-word entities is 15.70, while the EFG average value of the entire bigrams collection is 1.18.

Furthermore, we can see in Table 1 that the average function value is different for each entity length (i.e., average SCP value of 2-word entities is different from the average SCP value of 3-word entities). Therefore, three thresholds were determined for each function. All thresholds were experimentally evaluated.

5. ALGORITHM

We propose a two phase algorithm that uses entity based ranking to improve precision and recall in 'top results'. For the 'offline' phase, we assume global information is available. Here STAT is either SCP or EFG, and t_2, t_3, t_4 are the three relevant SCP or EFG thresholds.

Performed 'offline' (preprocessing):

ENTITY CORPUS GENERATION(STAT, t_2, t_3, t_4)

1. Collect metadata & n-gram statistics from descriptors
2. for every n-gram $2 \le n \le 4$ do
   i. value $\leftarrow$ STAT (n-gram)
   ii. if $value \ge t_n$
      - then add-entity(n-gram)
3. for each metadata term do (optional)
   i. add-entity(metadata)

Performed 'online':

RANKING (User-Query)

1. Parse Query as follows:
   i. Collect n-gram statistics based on Query terms
   ii. for $n \leftarrow 4$ to 2
      i. if n-gram appears in entity-corpus
         - store(n-gram)
         - remove any other n-gram $2 \le n \le 4$ that overlaps with the current one
2. Search P2P network for Query terms and get result-set
3. For each res $\in$ result-set
   i. if res contains all stored n-grams
      - then add res to top group
4. Rank top group based on group size (results in top group are always ranked the highest)

6. SEARCH RESULTS

Using the 1,000,000 file descriptors collection, plus three different randomly chosen sets of 96 queries also collected from IRWire [3], we simulated a centralized P2P search, in which global information is available, and all the nodes are visited in each search; query evaluation was done manually by us. We first retrieved files using standard conjunctive queries and ranked the results based on group size (used as a baseline). We then re-ranked the results using our proposed entity ranking algorithm (we compare 4 different functions for multiword named entity: SCP, SCP + metadata, EFG, and EFG + metadata). We used the following well known performance measures: Precision at top k files: (we chose $k = 10$ and $k = 20$), Binary Preference (bpref) (with R=5), and Mean Reciprocal Rank (MRR).

Table 1. SCP and EFG Average Values for Known Entities Versus the Entire Collection.

	SCP Average Value	EFG Average Value
2-Word Entities	0.15	15.70
All Bigrams	0.02	1.18
3-Word Entities	0.09	1600.57
All Trigrams	0.01	31.00
4-Word Entities	0.16	1668335.46
All 4-Grams	0.04	1292.92

For MRR, we consider the first relevant result as the desired file. Our results are presented in Table 2; we denote the combination of SCP and metadata by SCP+, EFG and metadata by EFG+, and conjunctive queries by CQ. It is shown that both SCP and EFG achieve a significant (P-value < 0.01, paired t-test) improvement over the standard conjunctive queries. Metadata were only helpful in a few queries where both SCP and EFG failed to identify entities; as a result, it only provides a minor improvement.

In this work we assumed a centralized P2P network; we leave the decentralized network study and scalability issues for future work.

Table 2. Standard CQ with Group Size Ranking Vs. CQ with SCP, SCP+, EFG, EFG+ Entity Ranking

	P@10	P@20	Bpref	MRR
CQ	0.60	0.54	0.65	0.61
SCP	0.64	0.59	0.68	0.62
SCP+	0.66	0.60	0.69	0.63
EFG	0.65	0.59	0.68	0.63
EFG+	0.66	0.61	0.69	0.63

7. REFERENCES

[1] da Silva J. F., and Lopes G. P. A local Maxima method and a Fair Dispersion Normalization for extracting multi-word units from corpora. In Sixth Meeting on Mathematics of Language, 1999.

[2] Merhav Y. and Frieder O., On Filtering Irrelevant Results in Peer-to-Peer Search, ACM 23rd Symposium on Applied Computing (SAC), Fortaleza, Brazil, 2008.

[3] Sharma S., Nguyen L. T., and Jia D. IR-Wire: A Research Tool for P2P Information Retrieval. In Proceeding ACM SIGIR Wkshp. 2006.