Semantic Overlay Networks for P2P Content Search

Michalis Vazirgiannis

(joint work with C. Doulkeridis, K. Noer vag – NTUN)

DB-NET (Research Group on Data & Web Mining)
Dept. of Informatics, Athens University EB,
GREECE

WWW: http://www.db-net.aueb.gr/

Previous work in Web related topics

• Temporal Link analysis [WAW2004] - MPI
• Web Personalization
  – Semantics for WP [KDD2003, TOIT2003]
  – Link analysis (local biased PR) [ICDM2005]
• Pagerank Prediction [ANAW 2006]
• P2P Web search [ECDL2006, WWW2006]
Motivation

- Centralized search engines - issues about their future applicability
  - Coverage and scalability
    - Decreasing coverage of the static Web – not to mention deep web
  - Freshness
  - Potential for Information manipulation
  - Cost of providing the service

- Promising potential:
  - Web search over P2P architectures


P2P content networks - Challenges

- Lack of global knowledge
- Lack of central coordination
- Content Ranking
- Churn
- Trust/Fame...
- …
Semantic Overlay Networks for Search

- **Semantic Overlay Networks (SONs):** peers with thematically similar content are logically grouped
- Assuming SONs, queries can be forwarded to most similar SONs
- **Benefits:**
  - Reduced query processing cost
  - Better results’ quality.
- **Challenges**
  - SON creation
  - Applicability to web search


Outline

- **Decentralized Semantic Overlay Network Generation [ECDL 2006]**
- **Semantic Overlays for Web Search [WWW2006]**
- Contributions - Future Work
DESENT - Preliminary Concepts

- Semi-structured hierarchical P2P architecture
- Unsupervised, decentralized and distributed SON generation
  - Phase 1: Local clustering
  - Phase 2: Zone initiator selection
  - Phase 3: Zone creation
  - Phase 4: Intra-zone clustering
  - Phase 5: Inter-zone clustering

Preliminary Concepts

- **Zones**: partitions of the P2P network
- **Initiators**: peers (members of the partitions) assigned the local coordinator role
- **Cluster Representatives**: peers representing thematically focused groups of peers within a zone
Phase 1: Local Clustering

- Documents represented by feature vectors \( F_i: \{(f_{ij}, w_{ij})\} \)
- Document clustering
  - Clusters represented by feature vectors
- Each peer provides a set of feature vectors representing its content (in terms of clusters)

Phase 2: Initiator Selection

- load-balancing: \( S_Z \) peers per zone
- Random initiator selection
- Peer \( P_i \) is initiator, if:
  \[(IP_{pi} + T) \mod S_Z = 0\]
- Select initiators uniformly spread over the network
Phase 3: Zone Creation

- Initiators send \textit{PROBE} messages (synchronization issues)
- Zones are established stepwise
- Neighboring initiators become familiar of each other
- If necessary, zone splitting is performed

Phase 4: Intra-zone Clustering

- Initiator sends \textit{FVecProbe} messages to all peers
- Peers send their feature vectors back
- Initiator performs clustering
- Cluster representatives $R_i$ are selected
- $CD_2=(C_2,F_2,(A_5,A_7,A_8,A_9,A_4))$
- $R_i$ inform peers in their cluster ($C_i, F_i, R_i$)
Phase 5: Inter-zone Clustering

• Recursively apply merging of zones to create super-zones,
• Level-i representatives know level-(i-1) representatives, to form a cluster hierarchy
• Terminates when only one initiator is left

Final Organization

• Hierarchy of peers
  – Each peer "knows its initiator
  – A level-i initiator knows its level-(i-1) initiators and the level-(i+1) initiator of the super zone
  – Each initiator knows all cluster representatives in its zones
• Hierarchy of clusters
  – Each peer knows the clusters it belongs to and the representatives of these clusters
  – A cluster representative knows the peers in its cluster, the representatives at one level below and the representative of the super-cluster
Searching in DESENT

1. Local query processing
2. Query forwarded to most similar cluster
   \[ \text{max}\{\text{sim}(Q, C_i)\} / Q \]
3. Query sent to one top-level initiator
   a. Most similar top-level cluster determined and query forwarded to representative. If necessary, backtracking is performed
   b. All similar top-level clusters determined and query forwarded to representatives (exhaustive search)

3a) achieves low query latency
3b) achieves higher recall, at the cost of extra messages

Feasibility Analysis

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Default Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>1 KB/s</td>
</tr>
<tr>
<td>L</td>
<td>4</td>
</tr>
<tr>
<td>(\text{max}^p)</td>
<td>(S_g/N_f)</td>
</tr>
<tr>
<td>(N_c^1)</td>
<td>10</td>
</tr>
<tr>
<td>(N_c^2)</td>
<td>100</td>
</tr>
<tr>
<td>(N_f)</td>
<td>&gt; (S_g/4)</td>
</tr>
<tr>
<td>(N_P)</td>
<td>100,000</td>
</tr>
<tr>
<td>(N_P)</td>
<td>20</td>
</tr>
<tr>
<td>(S_{CD})</td>
<td>&gt; 1.5S_P</td>
</tr>
<tr>
<td>(S_{FL})</td>
<td>200 bytes</td>
</tr>
<tr>
<td>(S_{Z})</td>
<td>60 seconds</td>
</tr>
<tr>
<td>(S_{P})</td>
<td>10 seconds</td>
</tr>
<tr>
<td>(S_{R})</td>
<td>2</td>
</tr>
</tbody>
</table>

Table 1. Parameters and default values used in the cost models
Creation Cost and Construction Time

- Creation cost: for Sz = 100, just above 100MB
- Construction time: for ta = 30sec, approx. 1 hour

Basic Simulation Setup

- Simulator implemented in Java, as a centralized process
- 2 setups: a) 8000 peers b) 20000 peers
- Random network topology (GT-ITM, SQUARE: dense topologies)
- Reuters-21578 collection
  - 8000 documents belong to 60 categories
  - 20000 documents
- Feature extraction (tokenization, stop-word removal, stemming, keep top-k features)
- HAC (similarity threshold Ts)
- Synthetic query workload, #keywords with mean=2, st.dev.=1, random terms with freq.>1%, avg.query results= 159 (8000)
- Baseline: the documents that contain all query keywords
Zone Partitioning

- \( N_p = 20,000 \) peers
- Random graph, avg. degree = 10
- \( S_z = 100 \)
- y-value shows the average number of zones with zone size equal to the x-value
- without zone partitioning 30% of the zones have sizes > \( S_z \), some twice as large

Clustering Quality

- Clustering quality relative to centralized clustering
  \( \frac{F_{\text{measure}_D}}{F_{\text{measure}_C}} \)
- Stable behavior with network size \( N_p \)
- SONs of high quality
Search Recall

- Compare DESENT recall to flooding recall using the same number of messages
- DESENT outperforms flooding
  - $x_{3-5}$ (GT-ITM)
  - $x_{10}$ (SQUARE)

Outline

- **DESENT:** Scalable Decentralized Semantic Overlay Network Generation
- **SOWES:** Semantic Overlays for Web Search
- Contributions - Future Work
SOWES

- Routing using cluster hierarchy is efficient, but
  - No load-balancing
  - Not suitable (in general) for P2P – some peers keep too much knowledge, so if they fail… 😊

- Solution
  - No cluster hierarchy (but keep the zone hierarchy!)
  - Create connections between clusters at merging
  - Use connections for intra-cluster routing
  - Use cluster representatives for inter-cluster routing

Algorithm

**Algorithm 1: Algorithm for creating links during cluster merging.**

1: **Input:** Cluster $C_1$, Cluster $C_2$, $d$
2: **Output:**
3: 
4: Peer $P_1$, $P_2$
5: if (merge($C_1$, $C_2$)) then
6:   for $i = 1$ to $d$ do
7:     $P_1 = \text{getLeastConnectedPeer}(C_1, i)$
8:     $P_2 = \text{getLeastConnectedPeer}(C_2, i)$
9:     connect($P_1$, $P_2$)
10: end for
11: end if
Searching in SOWES

• Querying peer $Q_P$
  – Intra-cluster routing using connections
  – Contact cluster representative $R$
  – Inter-cluster routing at cluster representative level

• Improvements
  – Caching: a) cluster descriptors b) query results
  – Shortest-Path trees

Overlay Structure

• 8,000 peers
• Connectivity of generated overlays ($L_d$) and cluster representative ($L_g$)
• Small number of created connections ($L_d$)
• For $d=4$ only $L_d=5$ connections
• More connections at cluster representative level ($L_g$)
Recall

- Compare SOWES recall to flooding recall with *same number* of messages
- Recall for the *best k* results
- SOWES outperforms naïve flooding approaches

**Total Message Cost**

- SOWES search cost
- *Minimum* denotes the number of messages required if global knowledge is assumed
SOWES Search Cost

SOWES search cost
- Routing at cluster representative level
- Routing within overlays

Latency

- Number of hops (i.e. messages) required to retrieve the first $k$ (5) results

- Hops vs. # first results (8000 peers)
Outline

• DESENT: Scalable Decentralized Semantic Overlay Network Generation [ECDL 2006]
• SOWES: Semantic Overlays for Web Search [WWW2006]
• Contributions - Future Work

Contributions - Future Work

• Decentralized & Distributed SON creation
• Good Clustering results
• Query routing for unstructured P2P content nets
• Salient search recall

• Potential for Web Search
  • Distributed document clustering
    – Distributed dimensionality reduction
    – Distributed Spectral methods for global feature space reconstruction
  • More semantics
  • Beyond plain keyword-based search (i.e. document based queries)
  • Ranking
  • Larger scale experiments
Thank you !

http://www.db-net.aueb.gr/