Semantic Overlay Networks for P2P Content Search

Michalis Vazirgiannis

(joint work with C. Doulkeridis, K. Noervag – 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

- Web Content analysis [VLDB-J 2004, TKDE-J. 2004][PKDD2005]
- 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

J. Li, B. Loo, J.M.Hellerstein, M. Kaashoek, D. Karger, and R. Morris. On the feasibility of peer-to-peer web indexing and search. In *Proceedings of the 2nd IPTPS'03*, 2003.

3

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

[1] A. Crespo and H. Garcia-Molina. Semantic Overlay Networks for P2P Systems. Technical report, Stanford University, 2002.

5

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)

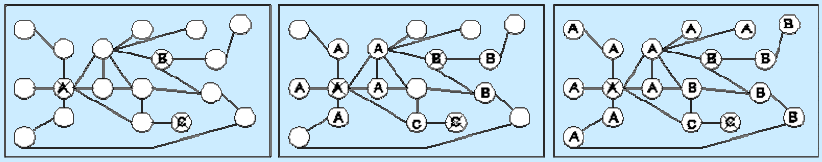
Phase 2: Initiator Selection

- load-balancing: S_Z peers per zone
- Random initiator selection
- Peer P_i is initiator, if:

  $(IP_{Pi} + T) \bmod S_Z = 0$
- Select initiators uniformly spread over the network

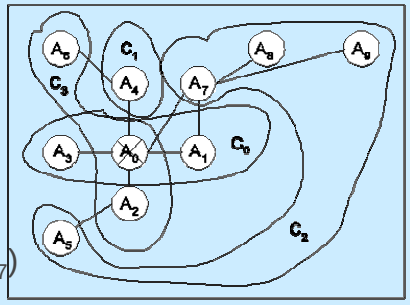
Phase 3: Zone Creation

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



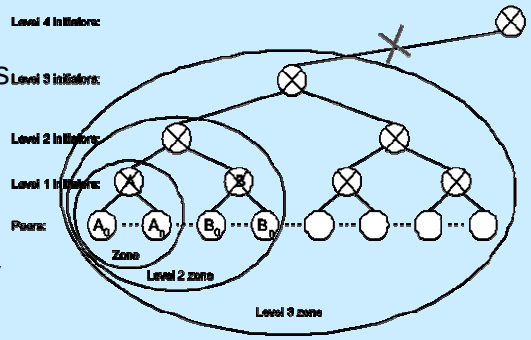
Phase 4: Intra-zone Clustering

- Initiator sends *FVecProbe* messages to all peers
- Peers send their feature vectors back
- Initiator performs clustering
- Cluster representatives R_i are selected
- $CD_2=(C_2,F_2,\{A_5,A_7,A_8,A_9\},A_7)$
- 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

Originating peer Q_P

1. Local query processing
2. Query forwarded to most similar cluster $\max\{sim(Q,C_i)\}$ / $Q_P \epsilon C_i$
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

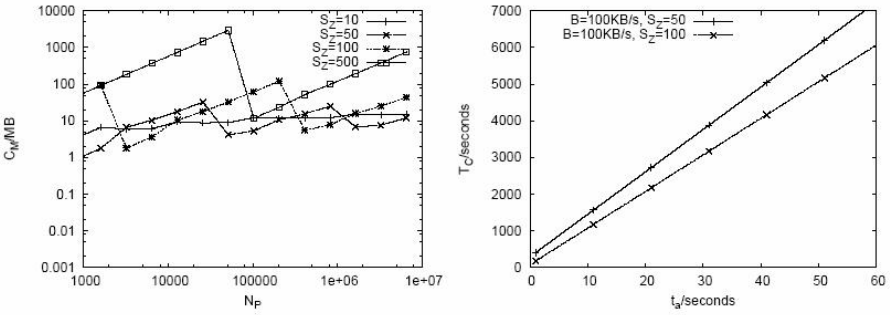
3b) achieves higher recall, at the cost of extra messages

Feasibility Analysis

	Parameter	Default Value
B	Minimum bandwidth available	1 KB/s
D_0	Avg. # of neighbors at level 0	4
D_i	Avg. # of neighbors at level i	S_Z
L	# of initiator levels	$\lfloor \log_{S_Z} N_P \rfloor$
min_F	Min. # of trees in top-level forest	$S_Z/4$
N_C^0	# of clusters per peer	10
N_C^i	# of clusters per level-i initiator	100
N_F	# of trees in top-level forest	$> S_Z/4$
N_i	# of peers/zones at level i	$\frac{N_P}{(S_Z)^i}$
N_P	Total # of peers in the network	1000000
r	Maximum number of hops between initiator and peer in a zone	20
S_{CD}	Size of a CD	$\approx 1.5 S_F$
S_F	Size of feature vector	200 bytes
S_M	Size of packet overhead[5]	60 bytes
S_Z	Avg. zone size	100
t_a	Time between synchronization points	60 seconds
v_c	# of time periods allowed for intra/inter-zone clustering step	10
v_m	Reduction value	2

Table 1. Parameters and default values used in the cost models.

Creation Cost and Construction Time



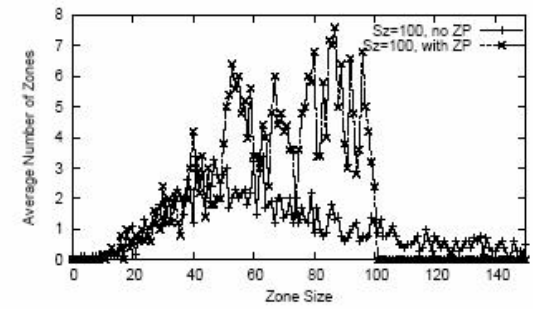
- Creation cost: for Sz = 100, just above 100MB
- Construction time: for ta = 30sec, approx. 1 hour

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

- Np = 20,000 peers
- Random graph, avg.degree=10
- Sz = 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 > Sz, some twice as large

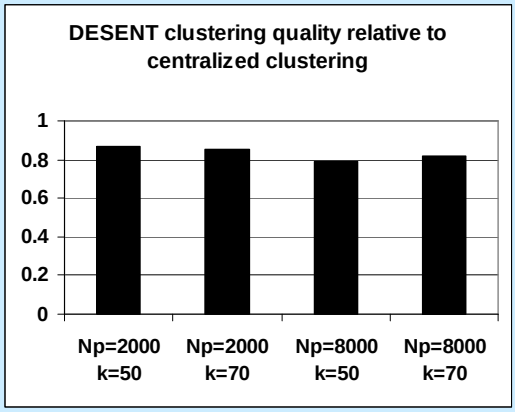


Clustering Quality

- Clustering quality relative to centralized clustering

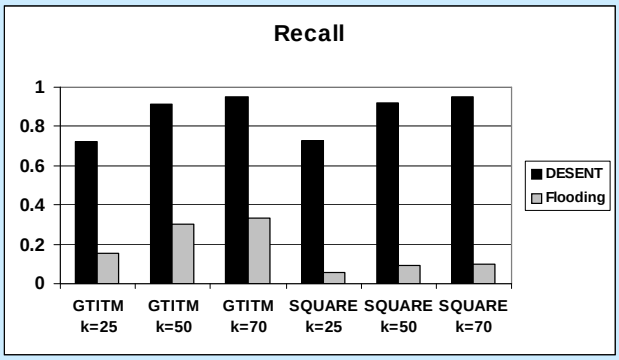
$F\text{-measure}_D / F\text{-measure}_C$

- Stable behavior with network size Np
- SONs of high quality



Search Recall

- Compare DESENT recall to flooding recall using the *same number* of messages
- DESENT outperforms flooding
  - x3-5 (GT-ITM)
  - x10 (SQUARE)

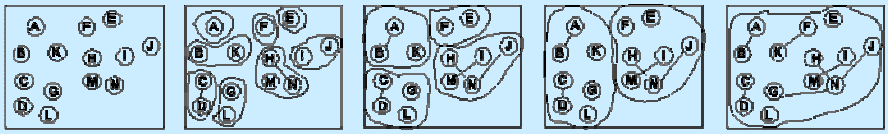


Outline

- 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.

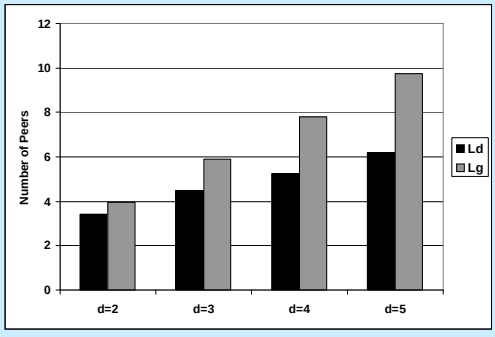
```
 1: Input: Cluster C_1, Cluster C_2, d
 2: Output:
 3:
 4: Peer P_1, P_2
 5: if (merge(C_1, C_2)) then
 6:    for i = 1 to d do
 7:       P_1 = getLeastConnectedPeer(C_1, i)
 8:       P_2 = getLeastConnectedPeer(C_2, i)
 9:       connect(P_1, P_2)
10:    end for
11: end if
```

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*

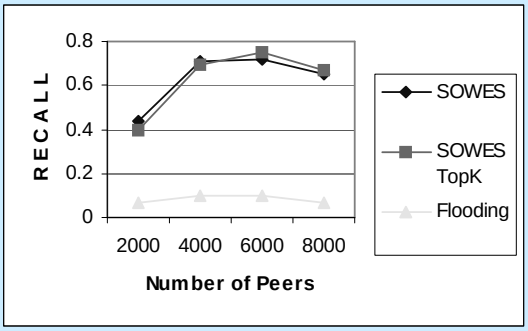
Overlay Structure

- 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)



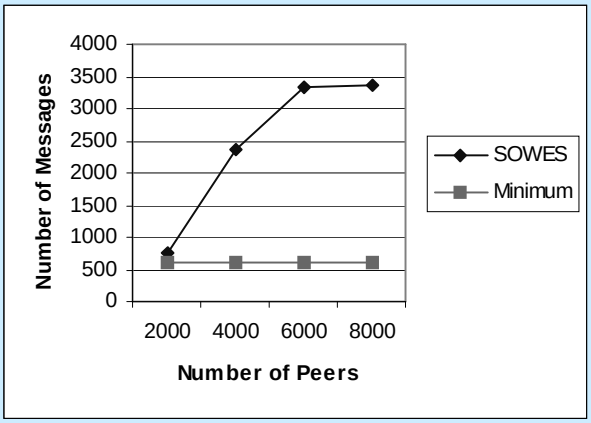
Recall

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



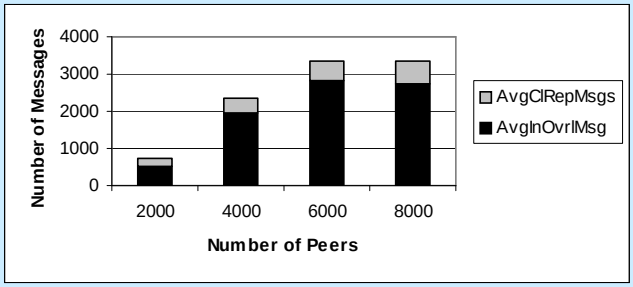
Total Message Cost

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



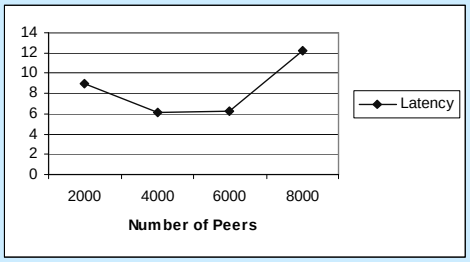
SOWES Search Cost

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

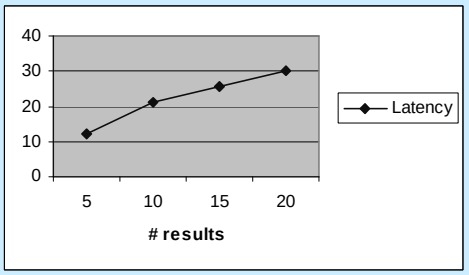


Latency

- 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/