Bloggers, Topics and Tags: An Analysis for a Blog Recommender System

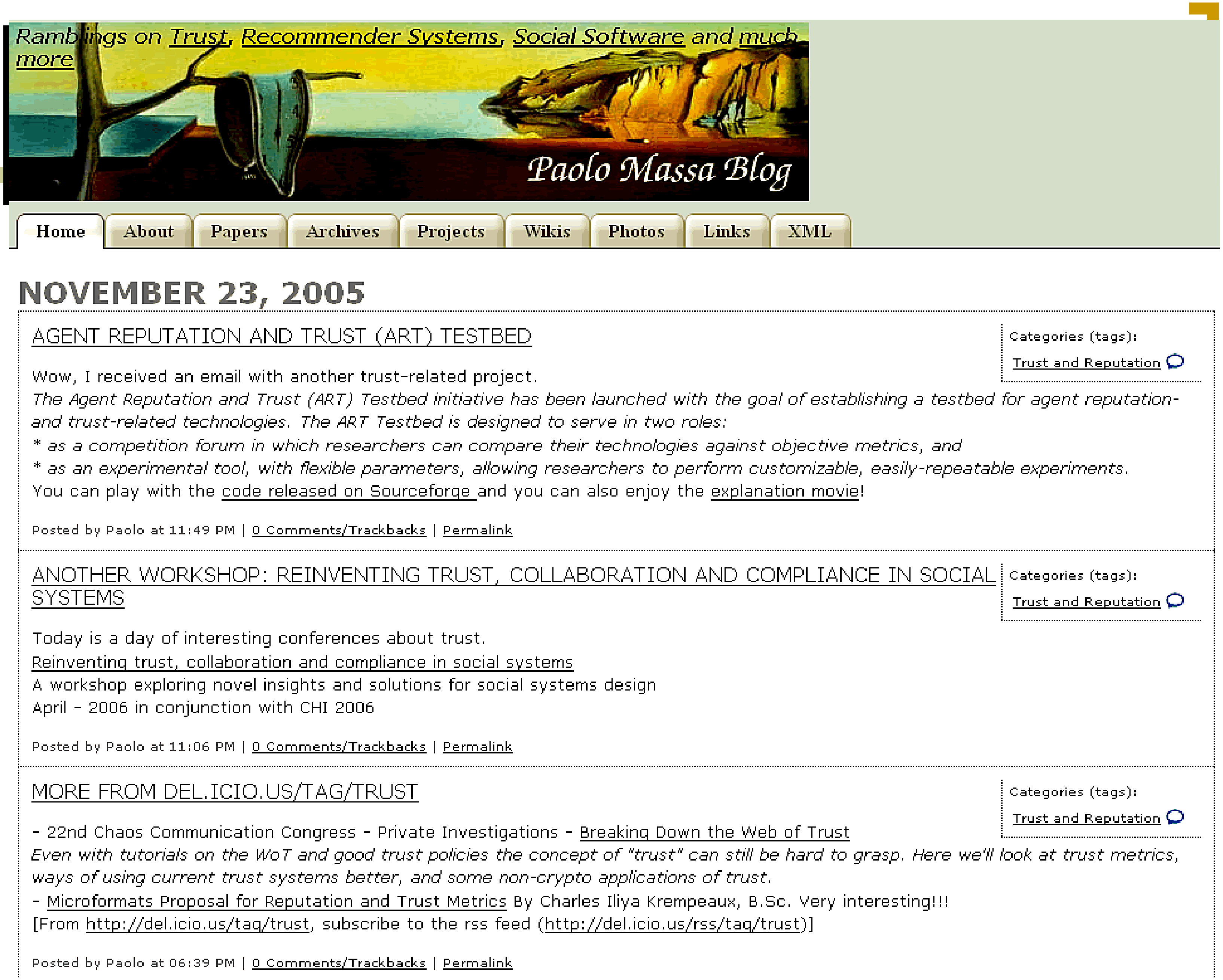
Conor Hayes
ITC-IRST
Trento,
Italy

Outline

- Context: Blog recommender system: present relevant topics and the best blog sources for a particular topic
- Motivations: Organizing the Blogosphere
    - Semantic Web
    - Web 2.0
    - Knowledge discovery
- Tagging
    - Tags as a primary partitioning strategy
    - Tags as a secondary strategy: Using tags to verify topic integrity
    - Using tags to identify potential topic authorities
        - Discover relevant blogs for topics uncovered using clustering
        - Discover bloggers that are consistent over time

Blogs

- Web site – journal style entries
  - Dated in reverse chronological order
  - Generally written by a single user
  - Regularly updated
- Distributed publishing
  - No technical, editorial constraints (apart from state censorship)
- Blogosphere: popular collective term for blogs viewed as a social network
- Exponential growth:
  - Technorati: 50 millions blogs (July 2006), doubling in size every 6 months
- Increasingly important indicator of public opinion on politics, technology, current affairs



Ramblings on Trust, Recommender Systems, Social Software and much more

Paolo Massa Blog

NOVEMBER 23, 2005

AGENT REPUTATION AND TRUST (ART) TESTBED

Categories (tags):
Trust and Reputation

Wow, I received an email with another trust-related project.
The Agent Reputation and Trust (ART) Testbed initiative has been launched with the goal of establishing a testbed for agent reputation- and trust-related technologies. The ART Testbed is designed to serve in two roles:
** as a competition forum in which researchers can compare their technologies against objective metrics, and*
** as an experimental tool, with flexible parameters, allowing researchers to perform customizable, easily-repeatable experiments.*
You can play with the code released on Sourceforge and you can also enjoy the explanation movie!

Posted by Paolo at 11:49 PM | 0 Comments/Trackbacks | Permalink

ANOTHER WORKSHOP: REINVENTING TRUST, COLLABORATION AND COMPLIANCE IN SOCIAL SYSTEMS

Categories (tags):
Trust and Reputation

Today is a day of interesting conferences about trust.
Reinventing trust, collaboration and compliance in social systems
A workshop exploring novel insights and solutions for social systems design
April - 2006 in conjunction with CHI 2006

Posted by Paolo at 11:06 PM | 0 Comments/Trackbacks | Permalink

MORE FROM DEL.ICIO.US/TAG/TRUST

Categories (tags):
Trust and Reputation

- 22nd Chaos Communication Congress - Private Investigations - Breaking Down the Web of Trust
Even with tutorials on the WoT and good trust policies the concept of "trust" can still be hard to grasp. Here we'll look at trust metrics, ways of using current trust systems better, and some non-crypto applications of trust.
- Microformats Proposal for Reputation and Trust Metrics By Charles Iliya Krempeaux, B.Sc. Very interesting!!!
[From http://del.icio.us/tag/trust, subscribe to the rss feed (http://del.icio.us/rss/tag/trust)]

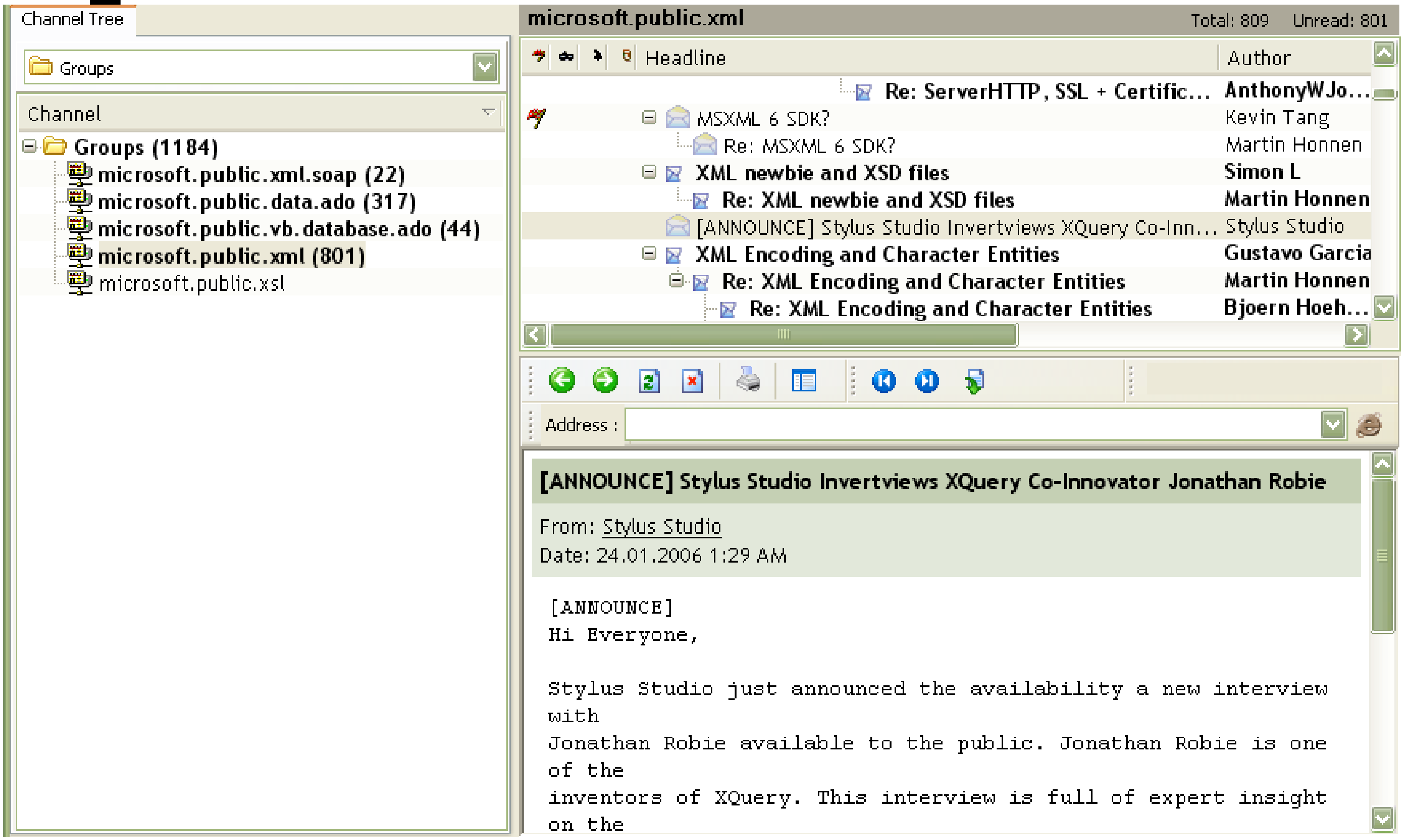
Posted by Paolo at 06:39 PM | 0 Comments/Trackbacks | Permalink

Blogs vs. Usenet

- Blogosphere is *user-centred*
  - Distributed architecture
  - Topic organisation is locally defined by tags
  - Not easy to find distributed posts related to the same topic
  - Not easy to identify authorities on a a particular topic

- Usenet is *topic-centred*
  - Logically centralised architecture
  - Topic organisation is *a priori* defined by newsgroup heading, and subject headers
  - Users know where to go to find information on a particular topic

Newsgroup organisation



Channel Tree

Groups

Channel

- Groups (1184)
  - microsoft.public.xml.soap (22)
  - microsoft.public.data.ado (317)
  - microsoft.public.vb.database.ado (44)
  - microsoft.public.xml (801)
  - microsoft.public.xsl

microsoft.public.xml — Total: 809 Unread: 801

Headline	Author
Re: ServerHTTP, SSL + Certific...	AnthonyWJo...
MSXML 6 SDK?	Kevin Tang
Re: MSXML 6 SDK?	Martin Honnen
XML newbie and XSD files	Simon L
Re: XML newbie and XSD files	Martin Honnen
[ANNOUNCE] Stylus Studio Invertviews XQuery Co-Inn...	Stylus Studio
XML Encoding and Character Entities	Gustavo Garcia
Re: XML Encoding and Character Entities	Martin Honnen
Re: XML Encoding and Character Entities	Bjoern Hoeh...

Address :

[ANNOUNCE] Stylus Studio Invertviews XQuery Co-Innovator Jonathan Robie

From: Stylus Studio
Date: 24.01.2006 1:29 AM

```
[ANNOUNCE]
Hi Everyone,

Stylus Studio just announced the availability a new interview
with
Jonathan Robie available to the public. Jonathan Robie is one
of the
inventors of XQuery. This interview is full of expert insight
on the
```

A Topic-Centred Blogosphere?

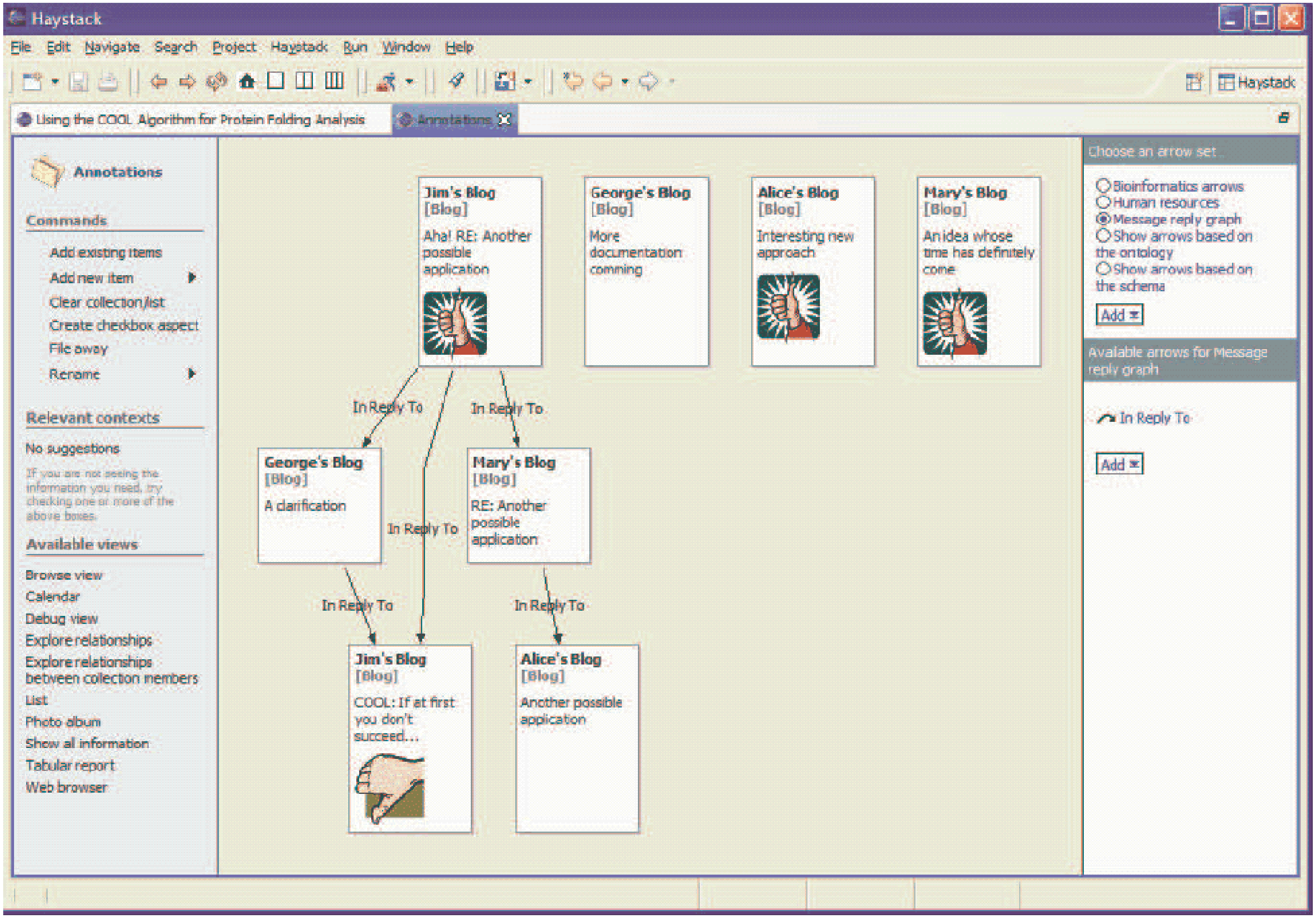
Top down approaches:

- **Semantic Web**
  - link blogs using SW standards, i.e, Semantic Blogging (Cayzer 2004), Haystack (Karger & Quan, 2005), SIOC (Bojars, Breslin, Harth, Decker 2005)
  - User Adoption?

Bottom up approaches:

- **Emergent classification: Tagging, Tag Clouds**
  - Tags: simple propositional entities, locally defined, easy-to-use
  - No global co-ordination on the relationship between tags
  - The problem of polysemes/synonyms
- **Knowledge discovery**
  - Clustering + online recommender systems
  - Link analysis: Majority of blogs have little or no inward connections

Haystack (Karger & Quan, 2005)



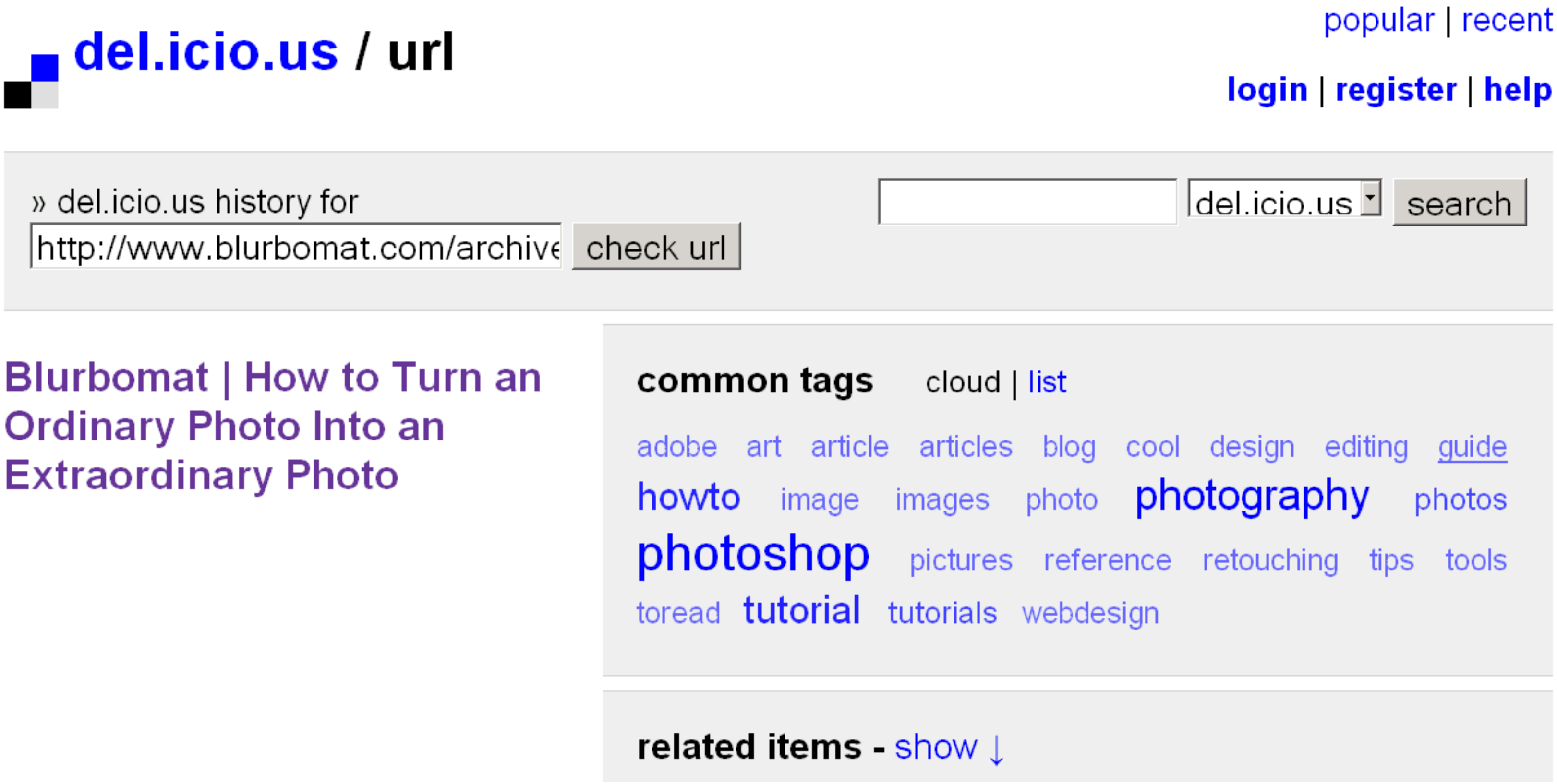
Semantic Web vision

- Developing a set of standard to add semantics to web content
- Perhaps, promised too much in the early days
- Ideal: RDF-based data model allows sophisticated, inference enabled querying of blogs
- User adoption problem
- Different degrees of SW ‘vision’
  - The SIOC project (Bojars, Breslin, Harth, Decker 2005) is a ‘bottom up’ attempt to bootstrap widespread use of SW standards
  - Export all information as RDF and allow applications to develop using this data
  - worry about ontology alignment, inferencing later
- RSS, FOAF : increasingly being adopted

Web 2.0

- Originally coined by Tim O’Reilly, O'Reilly Publishing
- Social networking, collaboration, architecture of participation…’
- New generation of applications that encourage online collaboration and sharing among users
- Web apps that ‘harness collective intelligence’:
  - Wikipedia, Technorati, Flickr, Del.icio.us
- Tagging = bottom up, local categorization
- Collaborative classification: ‘Tagsonomy’ not Taxonomy
- ‘Tagsonomy’ = ‘Tag Clouds’ = ‘Folksonomy’
- = collaborative, emergent categorization

Del.icio.us Tag Cloud



http://www.blurbomat.com/archive

Blurbomat | How to Turn an Ordinary Photo Into an Extraordinary Photo

common tags

adobe art article articles blog cool design editing guide howto image images photo photography photos photoshop pictures reference retouching tips tools toread tutorial tutorials webdesign

related items - show ↓

Tagging

Pros

- Lower categorization costs
- Represents the community’s perception of an Internet resource
- May quickly reflect changes in how a resource is categorized
- Increasingly widely used
- Works well quite well for sites like Del.icio.us where the multiple users tag a unique resource

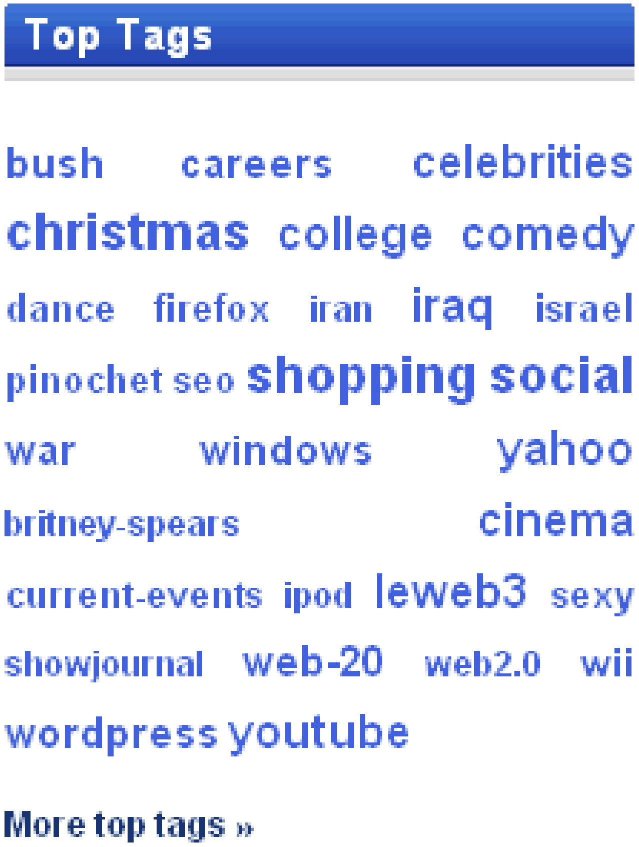
Cons

- No globally agreed list of tags
- No agreed ‘best practice’ for tagging
- No standard way of indicating tag equivalence
- No standard way of indicating relationships, tag hierarchies
- Sophisticated querying using tags not possible
- Semantics cannot be processed by machines
- Not as useful where the resource is tagged by a single person : **a blog post**

Technorati Tag Cloud

Blog tagging

- Bloggers only tag their own posts
- The tag describes a local ‘concept’ to which their post belongs
- Cannot check how other users have described this concept
- **Technorati** aggregates blog posts to produce a global tag cloud
- Less useful: not clear ..
  - what the relationship between tags are
  - what the emergent topics are

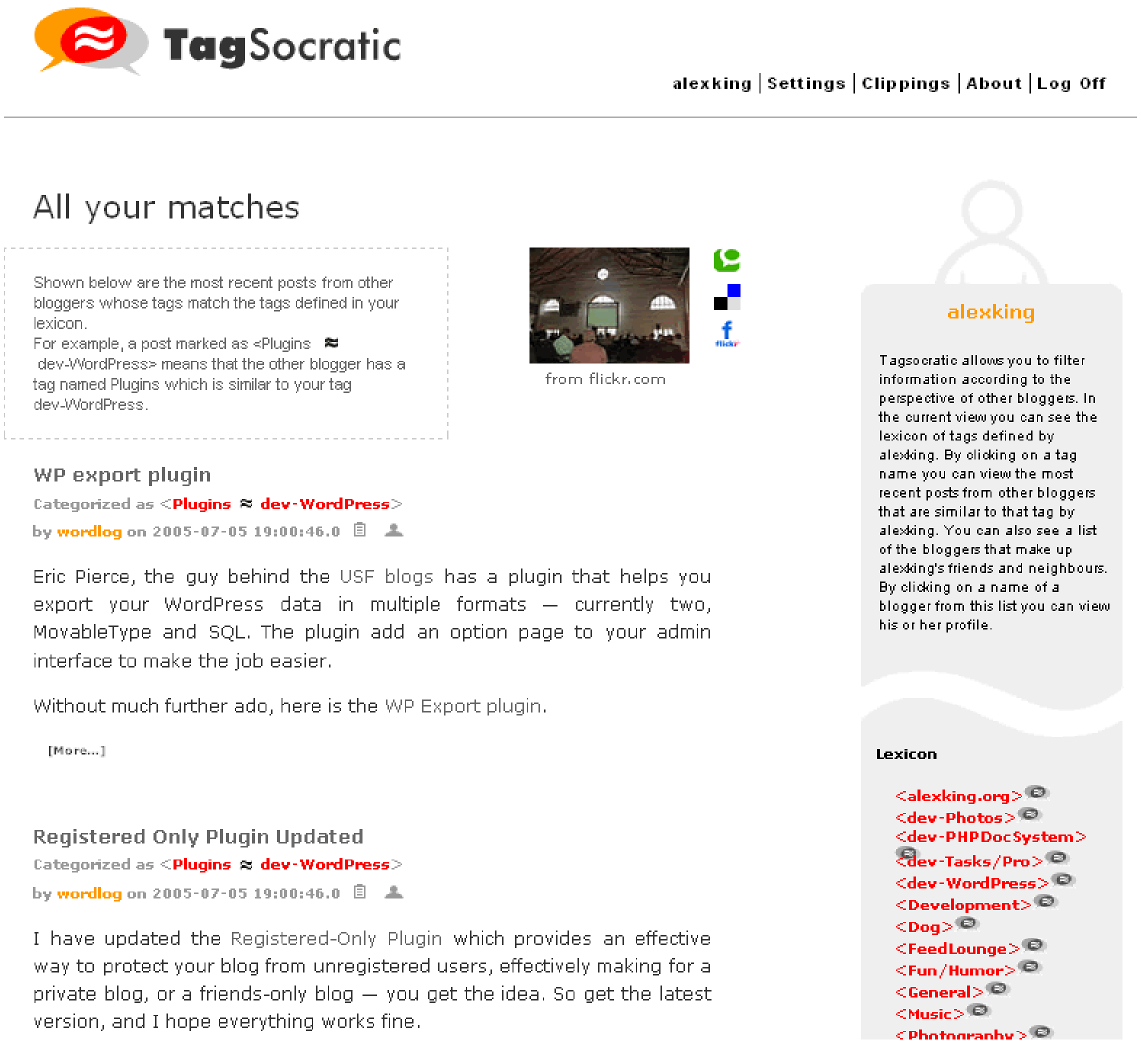


Top Tags

bush careers celebrities christmas college comedy dance firefox iran iraq israel pinochet seo shopping social war windows yahoo britney-spears cinema current-events ipod leweb3 sexy showjournal web-20 web2.0 wii wordpress youtube

More top tags »

Tagsocratic Blog Recommender



All your matches

Shown below are the most recent posts from other bloggers whose tags match the tags defined in your lexicon.
For example, a post marked as <Plugins ≈ dev-WordPress> means that the other blogger has a tag named Plugins which is similar to your tag dev-WordPress.

from flickr.com

WP export plugin

Categorized as <Plugins ≈ dev-WordPress>
by wordlog on 2005-07-05 19:00:46.0

Eric Pierce, the guy behind the USF blogs has a plugin that helps you export your WordPress data in multiple formats — currently two, MovableType and SQL. The plugin add an option page to your admin interface to make the job easier.

Without much further ado, here is the WP Export plugin.

[More...]

Registered Only Plugin Updated

Categorized as <Plugins ≈ dev-WordPress>
by wordlog on 2005-07-05 19:00:46.0

I have updated the Registered-Only Plugin which provides an effective way to protect your blog from unregistered users, effectively making for a private blog, or a friends-only blog — you get the idea. So get the latest version, and I hope everything works fine.

alexking

Tagsocratic allows you to filter information according to the perspective of other bloggers. In the current view you can see the lexicon of tags defined by alexking. By clicking on a tag name you can view the most recent posts from other bloggers that are similar to that tag by alexking. You can also see a list of the bloggers that make up alexking's friends and neighbours. By clicking on a name of a blogger from this list you can view his or her profile.

Lexicon

<alexking.org>
<dev-Photos>
<dev-PHPDocSystem>
<dev-Tasks/Pro>
<dev-WordPress>
<Development>
<Dog>
<FeedLounge>
<Fun/Humor>
<General>
<Music>
<Photography>

Partitioning ability of Tags

- How useful are tags at dividing a corpus of blogs into similar documents for retrieval purposes?
- How would tags compare to a clustering approach?
- Recall ability of tags ? How much of the corpus can they retrieve
- Alternatively, can tags be used in a supporting role to clustering?
  - Defining local topics
  - Verifying topic integrity
  - Identifying authoritative sources of information

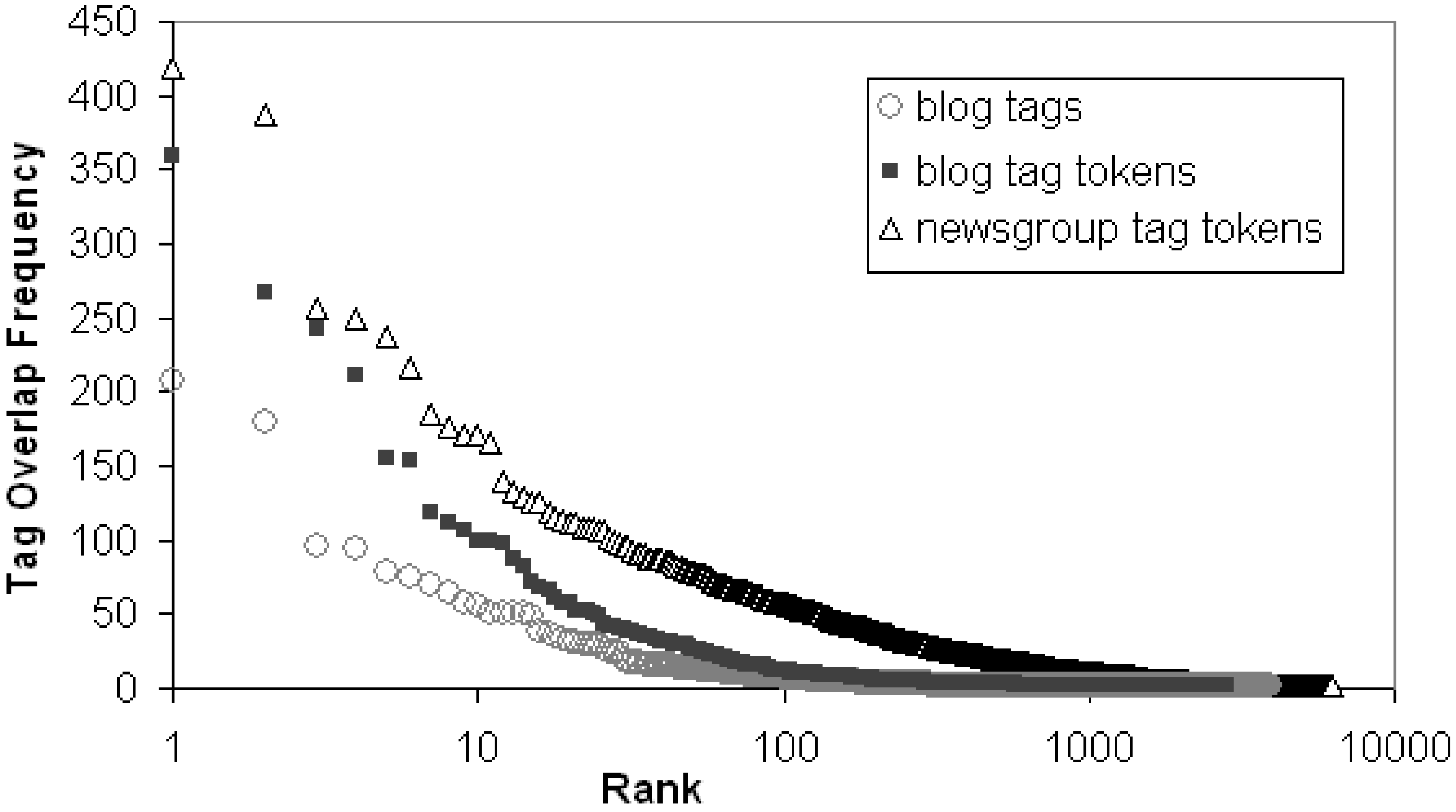
Recent Tagging analysis

- Brooks and Martinez (2005) analyzed the 350 **most used** tags in Technorati
- Tags do have some clustering ability: mean pair wise similarity for tagged documents was
  - higher than for documents grouped randomly
  - But lower for a selection of documents grouped by topic from Google
- Flat similarity distribution according to tag frequency
  - No significant difference in mean pairwise similarity between documents from frequently used and infrequently used tags
- We recorded similar results for the most popular tags in our data set….but

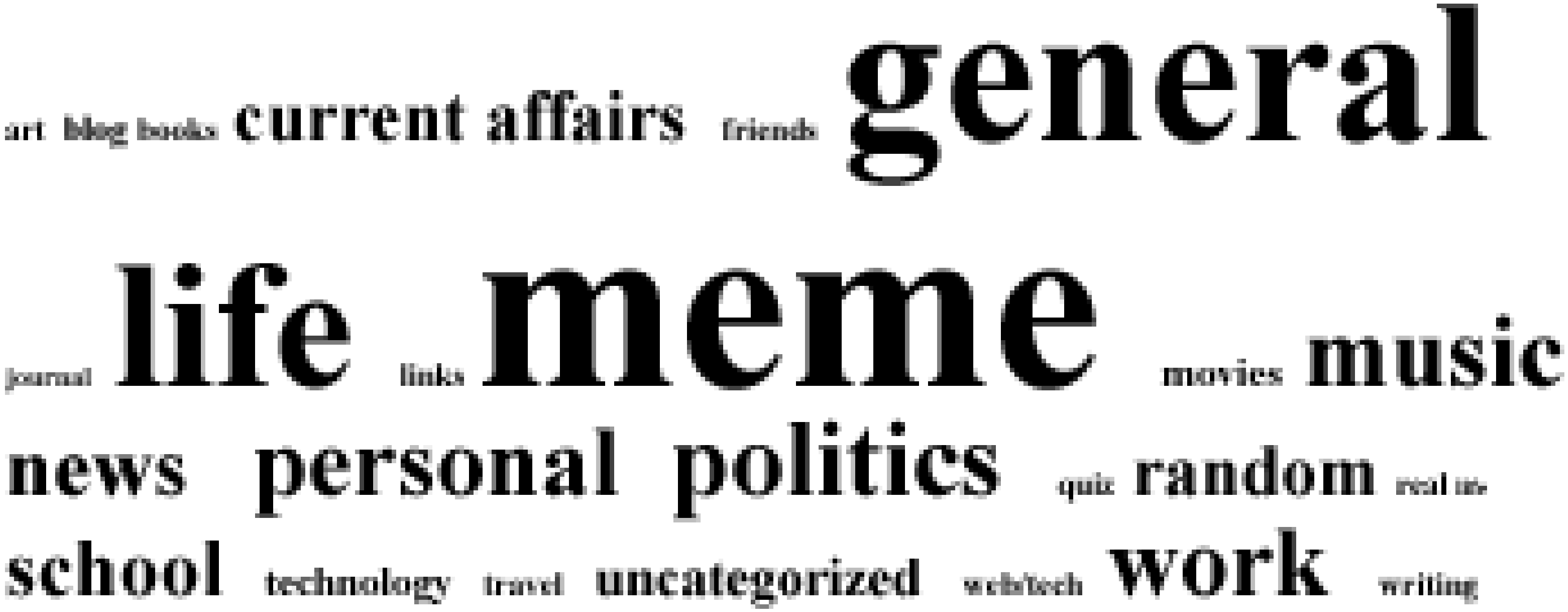
‘The Long Tail’

- Chris Anderson: Wired magazine 2004
- Clay Shirky: "Power Laws, Weblogs and Inequality"
  - A few blogs are highly linked to, the majority have very few or no links
- **The frequency distribution for tag overlap** also follows a power law
- Few very frequently used tags, very many tags used infrequently or just once
- For our data sets: 7209 blogs, 3934 tags
- 563 (14%) were used 2 or more times
- < less than 50% of blogs could be retrieved using tags
- So recall ability of unprocessed tags is quite poor

The Long Tail



Top 25 Tags in our global tag space

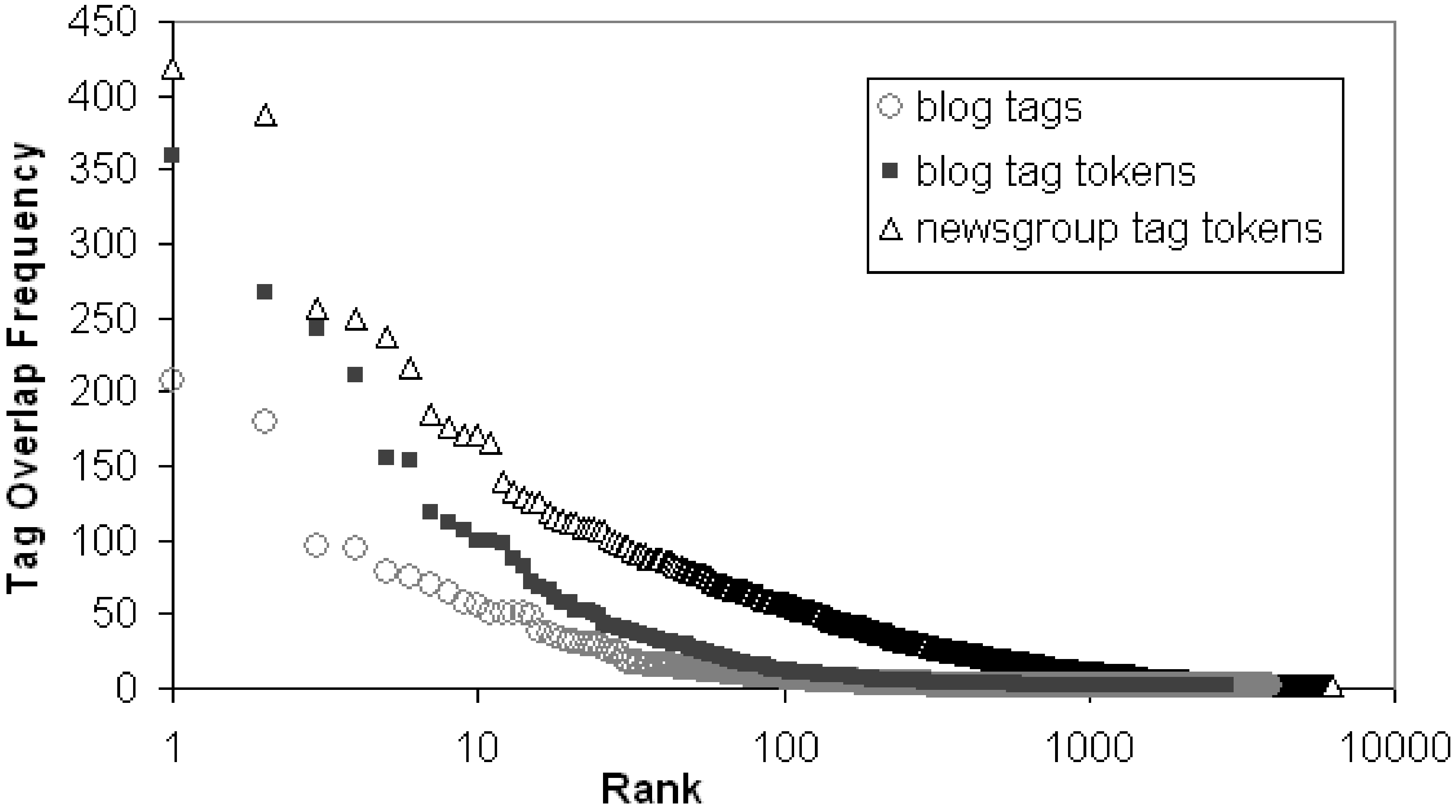


Tag token matching

- Our initial work compares the partitioning ability of tags against a standard clustering approach
    - Cluster using content
    - Cluster using tags
    - Compare results using cluster intra/inter distance measures

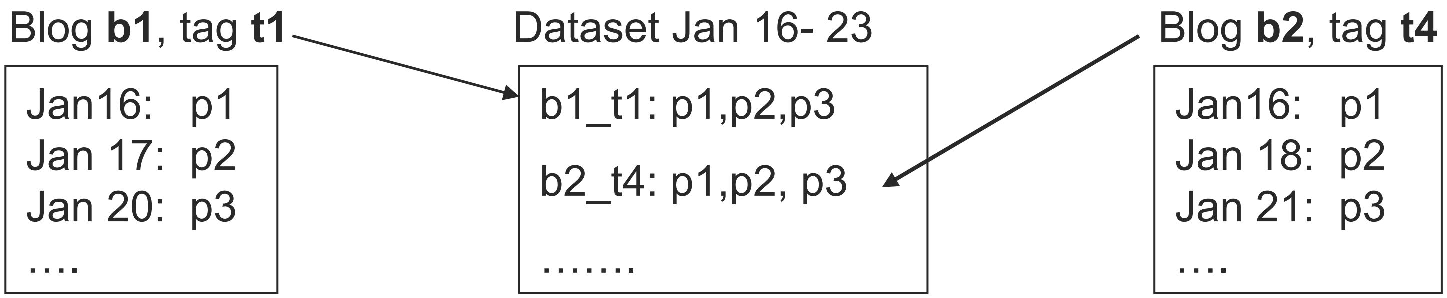
- To allow more frequent between tags, we used a standard IR technique
- Tokenizing, stemming and partial matching
    - “politics” doesn’t match “political news”
    - “polit” partially matches “polit new”
- Allowed us to make at least 1 match for 6064 of the 7209 blogs

The Long Tail



Our Blog Data

- We collected blog data from **Jan16 to Feb 27, 2006**
- We created 6 data sets, one for each week
- Each instance in each data set contains the posts from a **single tag**, from a **single blogger**



Blog **b1**, tag **t1**

Jan16: p1
Jan 17: p2
Jan 20: p3
….

Dataset Jan 16- 23

b1_t1: p1,p2,p3

b2_t4: p1,p2, p3

…….

Blog **b2**, tag **t4**

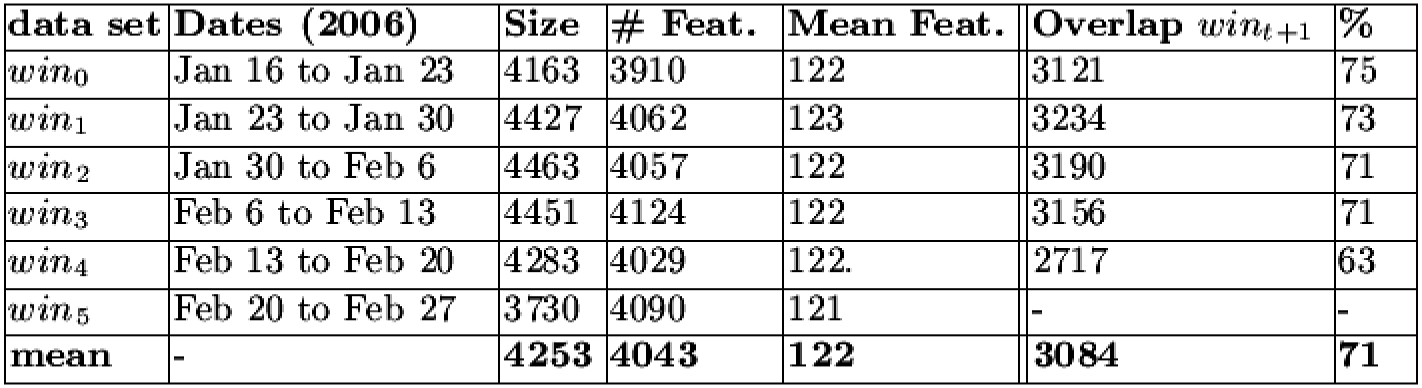
Jan16: p1
Jan 18: p2
Jan 21: p3
….

- An instance is included a data set **only if the user has posted in that week**
- Each instance contains posts from the data set week **plus** previous two weeks
- **E.g. if a blog is updated in week 3, the instance contains posts from weeks 3, 2 & 1.**

Processing the Data

For each data set:

- Removed stop words, stemmed and removed too frequent and infrequent words
- Removed docs with less than 15 tokens
- TF/IDF weighting, L^2 normalised



data set	Dates (2006)	Size	# Feat.	Mean Feat.	Overlap win_{t+1}	%
win_0	Jan 16 to Jan 23	4163	3910	122	3121	75
win_1	Jan 23 to Jan 30	4427	4062	123	3234	73
win_2	Jan 30 to Feb 6	4463	4057	122	3190	71
win_3	Feb 6 to Feb 13	4451	4124	122	3156	71
win_4	Feb 13 to Feb 20	4283	4029	122.	2717	63
win_5	Feb 20 to Feb 27	3730	4090	121	-	-
mean	-	**4253**	**4043**	**122**	**3084**	**71**

Table 1. The periods used for the windowed blog data set. Each period is from midnight to midnight exclusive. User overlap refers to the overlap with the same users in the data set for the next window.

Newsgroup data set

- For comparison purposes: well known labelled data set in text classification clustering
- 7183 users (similar to the # blog data set)
- posts to 20 newsgroup topics (class labels) over a 4 week period
- Each instance in our data set contains the posts from a single user
- No tags! Instead each post has a subject header
- For each user instance we synthesise a ‘tag’ from his subject header tokens
- Synthesised tag are longer than blog tags
  - Mean 5.5 tokens vs. mean 1.27 tokens

Clustering

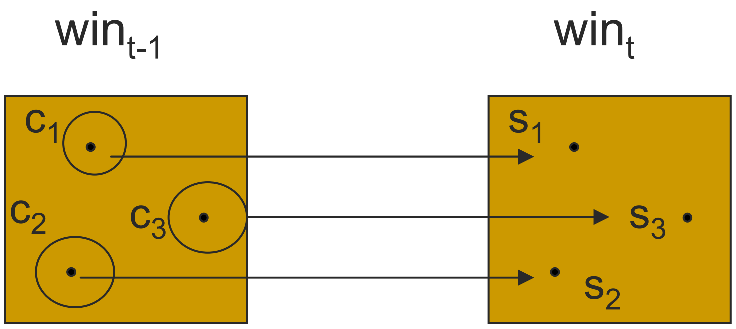
- **Goals:**
  1. Uncover latent structures or topics in the collection
  2. provide a means of summarisation or labelling
- **Spherical *k*-means :** (Dhillon et al. 2001)
  1. scales well to large collections
  2. produces interpretable concept centroids
- **Clustering quality:**
  - points in the same cluster should be close together;
  - points in different clusters should be far apart.

$$\mathcal{H}_r = \frac{\mathcal{I}_r}{\mathcal{E}_r} = \frac{\frac{1}{|S_r|}\sum_{d_i \in S_r} \cos(d_i, C_r)}{\cos(C_r, C)}$$

H_r: ratio of intra- to inter-cluster similarity (Zhao & Karypis 2004)

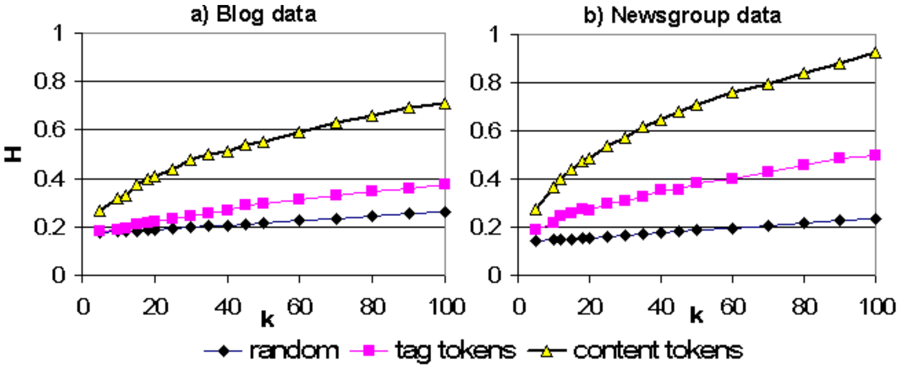
Clustering Windows

- We cluster each data set **in date order** at different *k*
- For the first data set, ***win_0***, seeds are chosen to maximise intra-seed distance
- For clustering win_0 to win_5
  - Seeds for win_t are based on the **final centroids** of clusters produced for win_{t-1}
  - provides continuity between clusterings on each data set

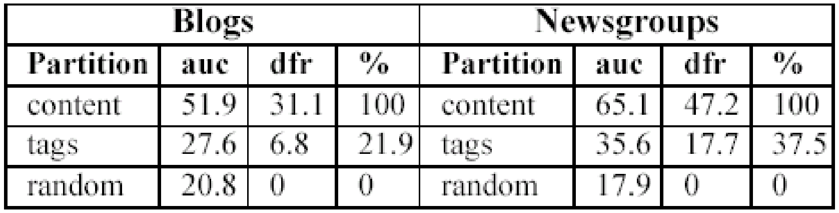


Partitioning by tag tokens

- Text clustering approach
- We compare clustering approach using tag tokens, content only and random clustering



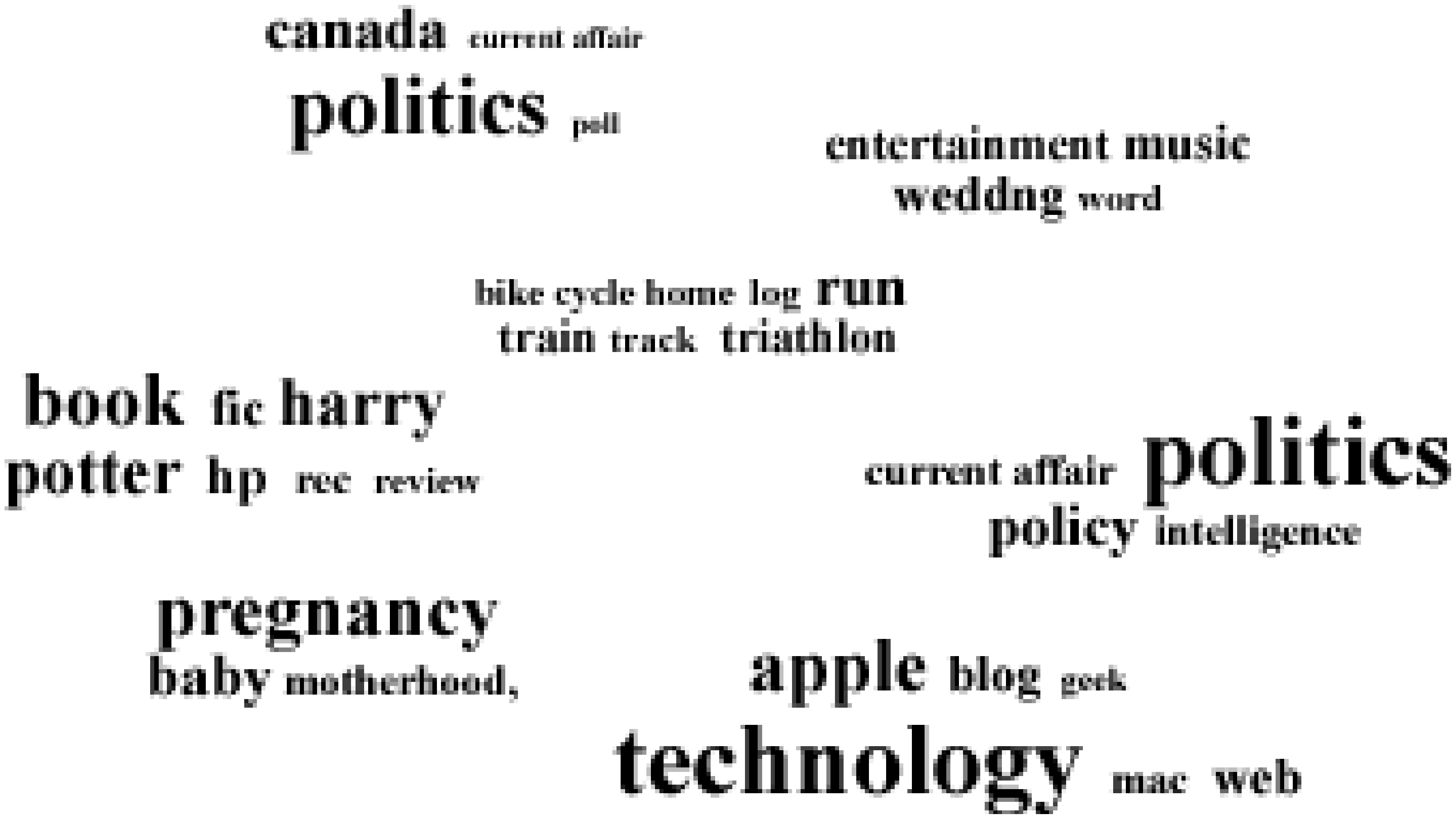
- Clustering by blog tags = slightly better than random a selection
- Newsgroup 'tags' perform much better
- Users adopt the subject headers of previous post when replying ?



Blogs				Newsgroups			
Partition	**auc**	**dfr**	**%**	**Partition**	**auc**	**dfr**	**%**
content	51.9	31.1	100	content	65.1	47.2	100
tags	27.6	6.8	21.9	tags	35.6	17.7	37.5
random	20.8	0	0	random	17.9	0	0

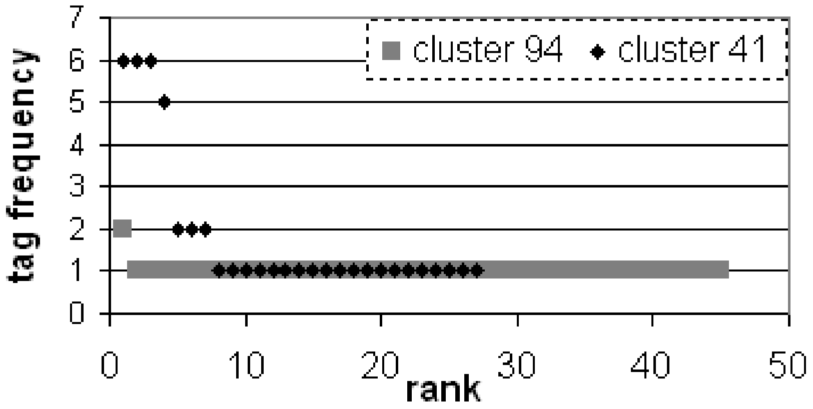
Table 2: A summary of the information in Figure 2. auc = area under curve. dfr = difference from random.

Clustering by content…. fragments the global tag space



Tag distribution per cluster

- Examining the tag token distribution **per cluster**
- We find a power law frequency distribution
- **Frequency distribution varies with cluster strength**



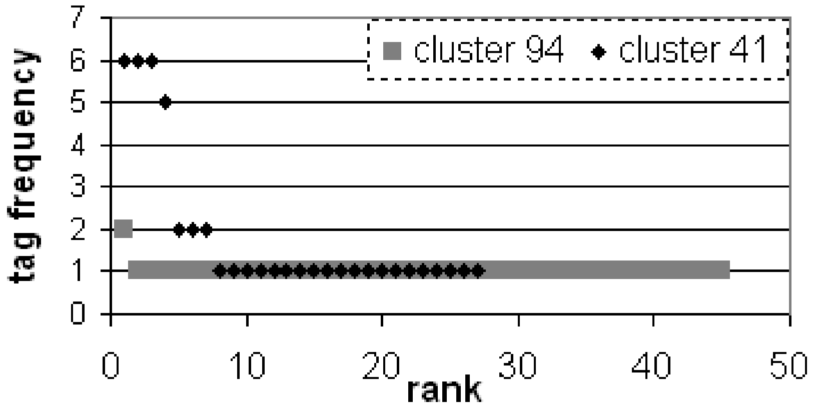
Example

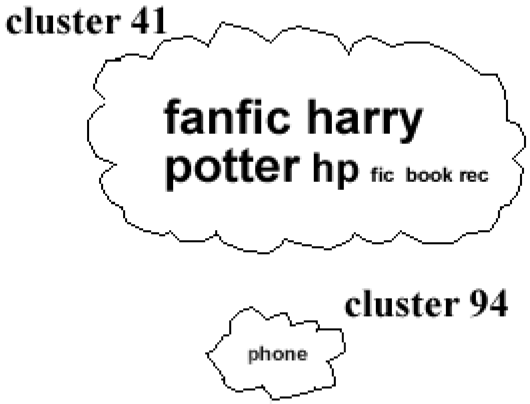
- Cluster 94 : low H
- Cluster 41 : high H

Relationship between cluster strength and tag distribution?

Intuition: probability of distributed users independently using the same tag token is higher in a well defined cluster topic than in cluster where the topic is weakly defined

Tag distribution per cluster





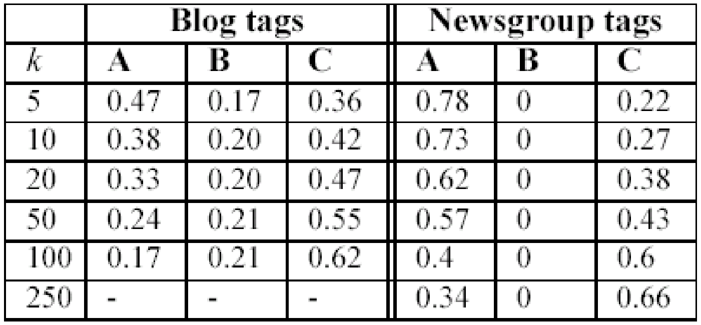
Strong clusters have large proportions of high frequency tags

Tag types

We can qualify the tag frequencies per cluster

- **C-tags** : tag tokens not repeated by any other user in the cluster
- **B-tags** : tokens where freq ≥ 2 and that occur in a 'large number' of clusters.

  examples : 'general', 'stuff', 'uncategorized', 'meme'
- **A-tags**: remaining tags. Freq ≥ 2



	Blog tags			Newsgroup tags		
k	**A**	**B**	**C**	**A**	**B**	**C**
5	0.47	0.17	0.36	0.78	0	0.22
10	0.38	0.20	0.42	0.73	0	0.27
20	0.33	0.20	0.47	0.62	0	0.38
50	0.24	0.21	0.55	0.57	0	0.43
100	0.17	0.21	0.62	0.4	0	0.6
250	-	-	-	0.34	0	0.66

Table 4: The table gives the mean fractions of a-tags, b-tags and c-tags per cluster at different values of *k*. The means are measured over the 6 windowed blog datasets.

A-tags represent the **tag token cloud** for a particular cluster

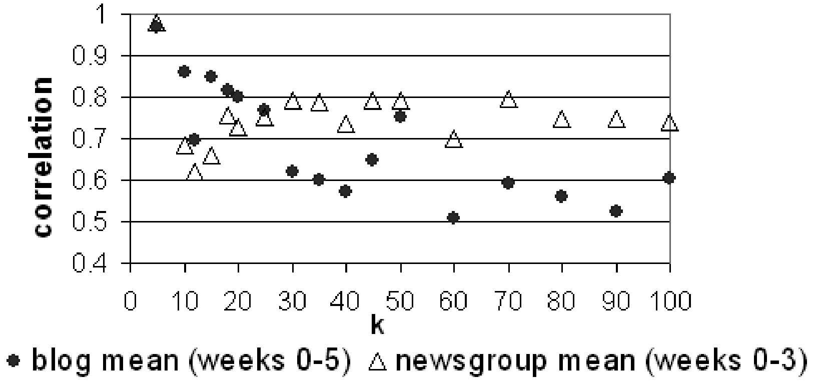
The T score

- The T score is the fraction of a-tags in a cluster, scaled by the cluster size

$$T_r = \frac{1}{|n|}\frac{\sum_{i=1}^{|A_r|}|a_i|}{\sum_{i=1}^{|A_r|}|a_i| + \sum_{i=1}^{|B_r|}|b_i| + \sum_{i=1}^{|C_r|}|c_i|}$$

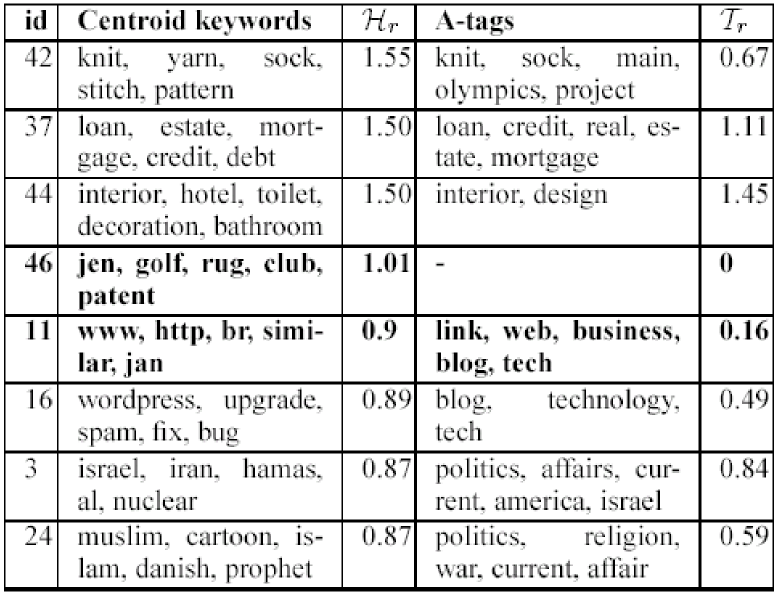
- A_r, B_r and C_r are the (disjoint) sets of a-, b- and c-tags in cluster r.
  n = size of cluster r

- **T_r and H_r strongly correlated**



What can we do with the T score?

- **Verify cluster integrity**
- Clustering using content tokens alone often produces **meaningless clusters** due to noise in the feature set
- Typical clustering evaluation measures cannot detect these – only humans can
- The T score can help
- A high H score **and** a low T score generally indicates a meaningless cluster



id	Centroid keywords	$\mathcal{H}_r$	A-tags	$\mathcal{T}_r$
42	knit, yarn, sock, stitch, pattern	1.55	knit, sock, main, olympics, project	0.67
37	loan, estate, mortgage, credit, debt	1.50	loan, credit, real, estate, mortgage	1.11
44	interior, hotel, toilet, decoration, bathroom	1.50	interior, design	1.45
46	**jen, golf, rug, club, patent**	**1.01**	-	**0**
11	**www, http, br, similar, jan**	**0.9**	**link, web, business, blog, tech**	**0.16**
16	wordpress, upgrade, spam, fix, bug	0.89	blog, technology, tech	0.49
3	israel, iran, hamas, al, nuclear	0.87	politics, affairs, current, america, israel	0.84
24	muslim, cartoon, islam, danish, prophet	0.87	politics, religion, war, current, affair	0.59

Table 3: The top 8 clusters in terms of $\mathcal{H}_r$ scores

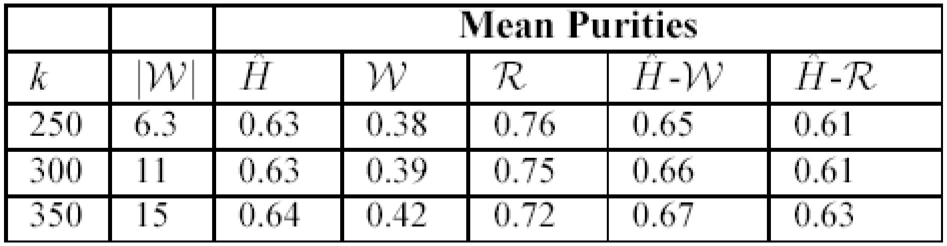
Evaluation

- We cluster the **labelled** newsgroup dataset
- For each instance we use a single tag token
- We use the T score to automatically identify weak clusters
- **Purity** P_r = fraction of the cluster made of instances of a single class (single newsgroup)

- We measure the mean purity of the top 20% of clusters = H’
- We remove clusters the identified by T = W
- measure the mean purity of H’ – W
- An increase means we have removed impure clusters

Evaluation 2

- Experiment conducted at 150 times; 50 times each at *k* = 250, 300, 350 using different seeds
- An increase in purity was observed in all 150 tests
- Difference significant at the 0.05 alpha level



		Mean Purities				
k	$\lvert\mathcal{W}\rvert$	$\hat{H}$	$\mathcal{W}$	$\mathcal{R}$	$\hat{H}$-$\mathcal{W}$	$\hat{H}$-$\mathcal{R}$
250	6.3	0.63	0.38	0.76	0.65	0.61
300	11	0.63	0.39	0.75	0.66	0.61
350	15	0.64	0.42	0.72	0.67	0.63

Table 6: The mean purities for each value of *k*.

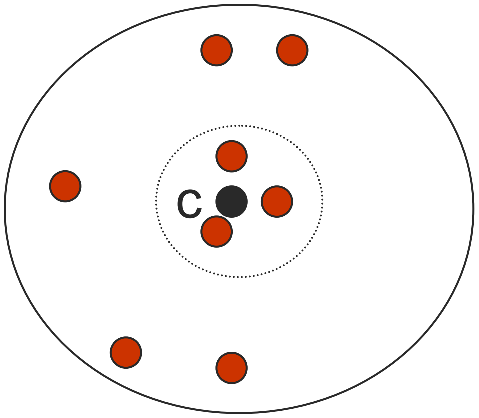
- The key result of this experiment is that $\mathbf{T_r}$ **allowed us to identify a subset of weak clusters, which the standard** $\mathbf{H_r}$ **score could not do**

To Recap

- A power law frequency distribution exists for global tag usage
- Majority of tags cannot be matched
- Partial matching on tokenized stemmed tags improves matching capabilities
- Using tag tokens to cluster document proves to be quite poor
- However, clustering by content allows us to subdivide the global tag space creating multiple local tag clouds
- Each tag cloud establishes a local relationship between tags
- Power law frequency distribution exists for tag usage within each cluster
- T score defines the relative strength of the high frequency section of the power law distribution
- T score allows us to identify semantically weak clusters that cannot be identified by other means
- Tags provide a useful supporting role to clustering

Identifying authorities 1

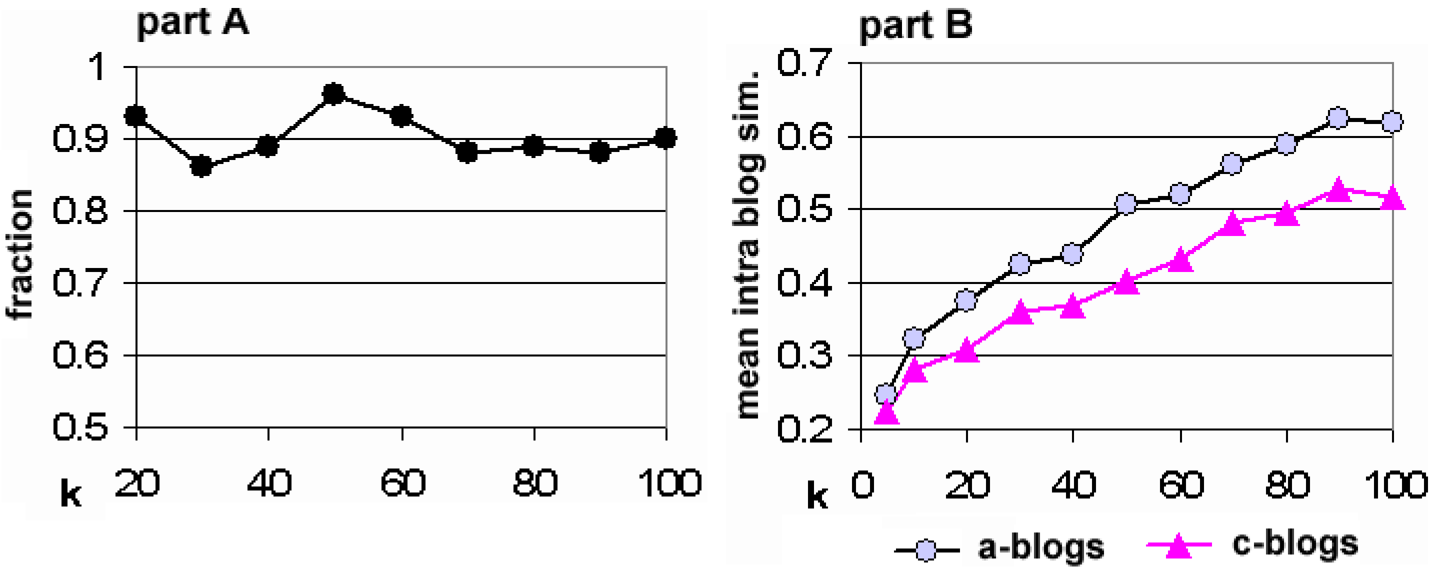
- How can we discriminate between ‘authoritative’ blogs and ‘shallow’ blogs : a relevance problem
- For each cluster topic, which blogs are the most useful to recommend?
- In each cluster, blogs have differing degrees of similarity to the topic defined by the cluster centroid
- blogs closer to the cluster centroid are more likely to be about the topic – possibly more informative



Hypothesis: blogs that contribute A-tag tokens are likely to be closer to the centroid

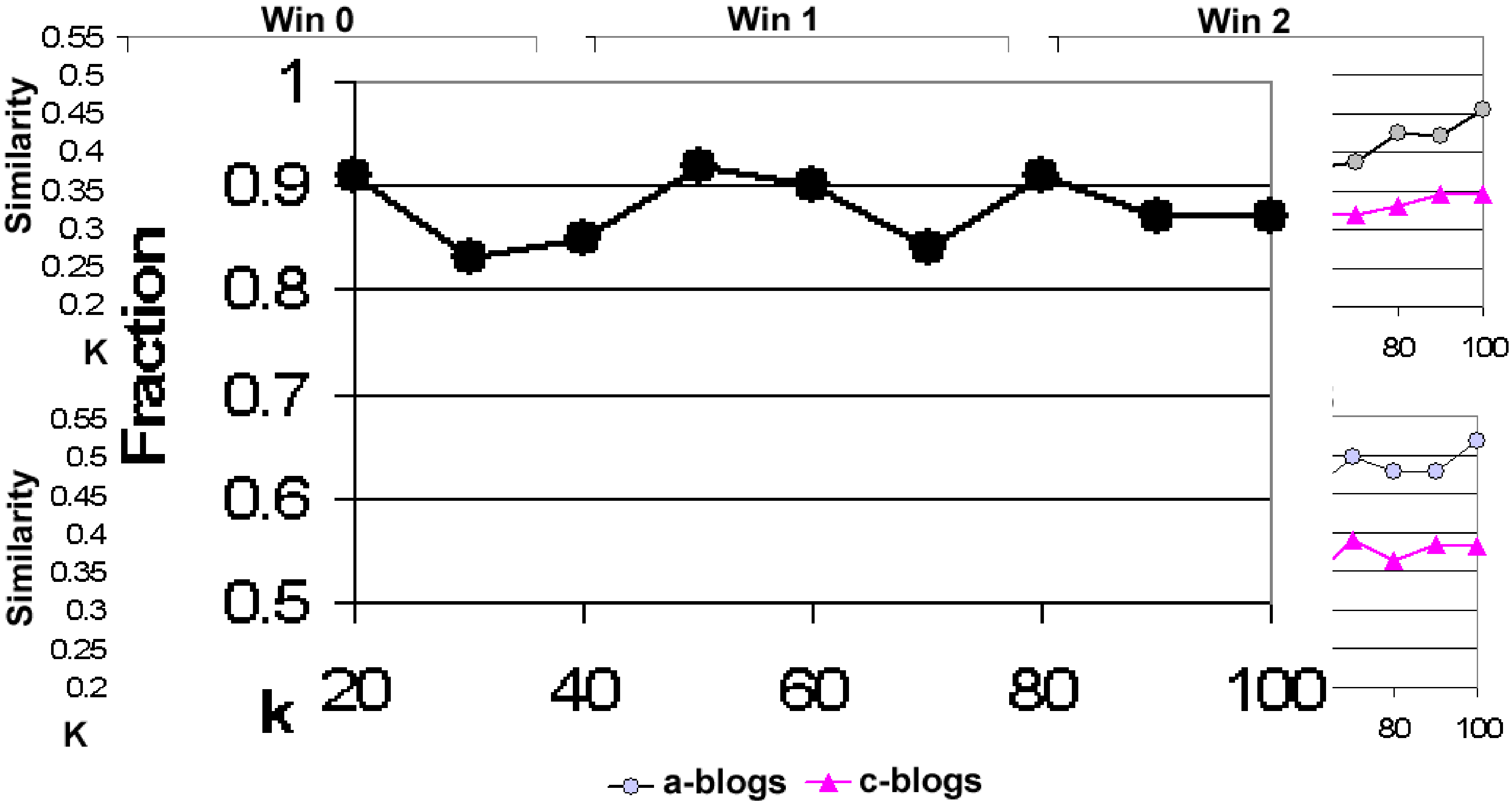
Why?

A- vs. C-blogs Intrablog similarity



- A-tag blogs are ‘tighter’ – more similar to each other than C-tag blogs

A- vs. C-blogs similarity to cluster centroid



- A-tag blogs tend to be more similar to the cluster centroid

Relevance?

- Within each cluster a-blogs form a tight subgroup which tends to be very similar to the cluster centroid
- Are they more *relevant* to the cluster concept than c-blogs?
- Van Rijsbergen’s cluster hypothesis suggests that similar docs are likely to more relevant to an information requirement than less similar documents
- The information requirement? = the cluster summary
- In application terms, the goal is to present to the user the most relevant blogs to the cluster summary
- How do we test relevance? The Google Oracle

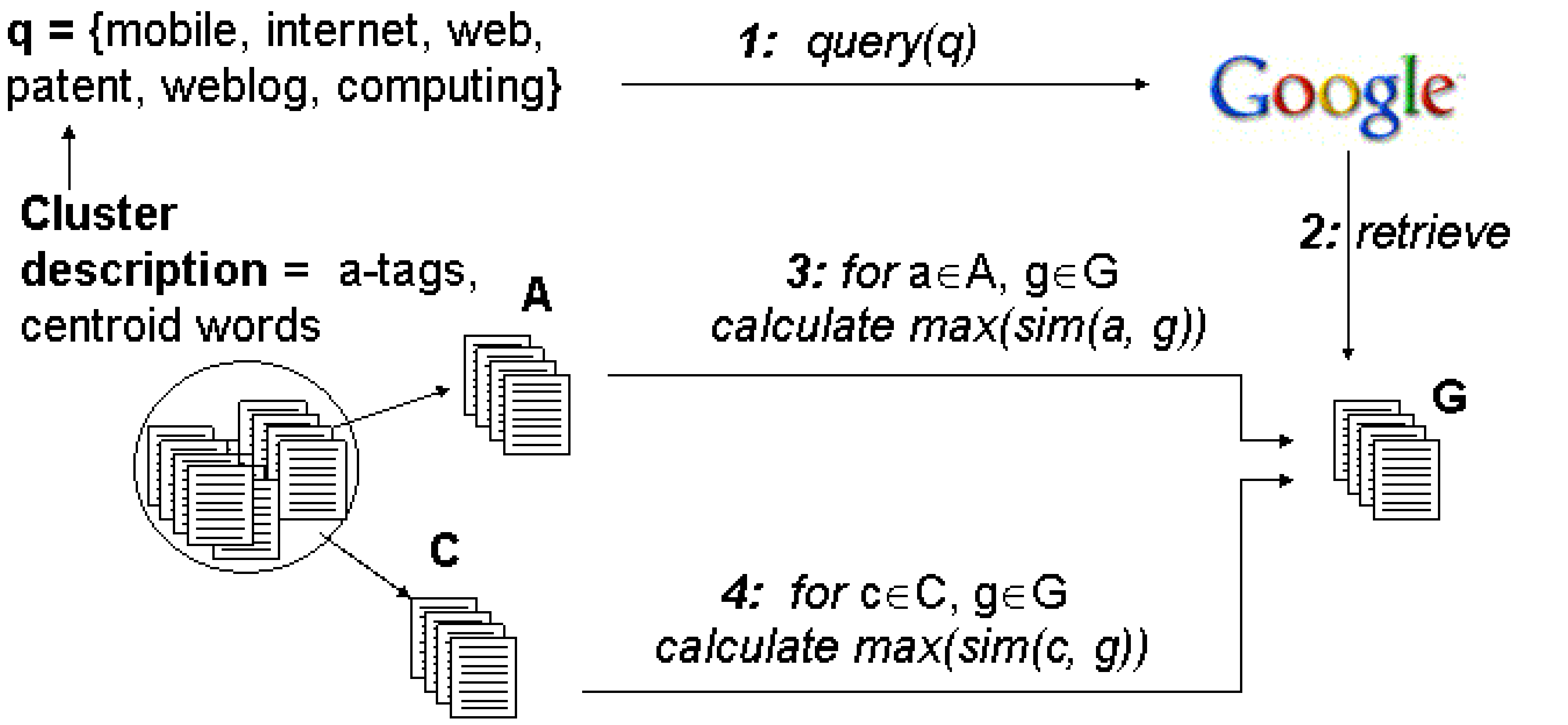
Verification 1: by Google

- In this experiment we test whether a-tag blogs are closer than c-tags to an independent assessment of the cluster topic – the **Google Oracle**

For each data set win_0, win_1, win_2, win_3, win_4, win_5

1. For each cluster: we submit the top 5 centroid key words as a Google query
2. We retrieve the 10 pages ranked by Google
3. For each a-tag blog - we measure its similarity to each page: record **max similarity**
4. For each c-tag blog - we measure its similarity to each Google page: **record max similarity**
5. Record means for each cluster
6. Repeat 1–5 using a-tag descriptions

Verification 2: by Google



q = {mobile, internet, web, patent, weblog, computing}

Cluster description = a-tags, centroid words

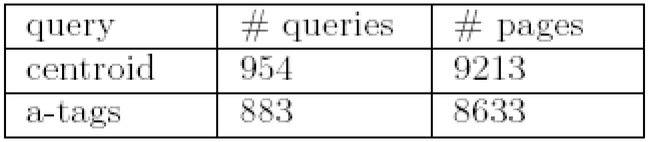
1: query(q)

2: retrieve

3: for a∈A, g∈G *calculate max(sim(a, g))*

4: for c∈C, g∈G *calculate max(sim(c, g))*

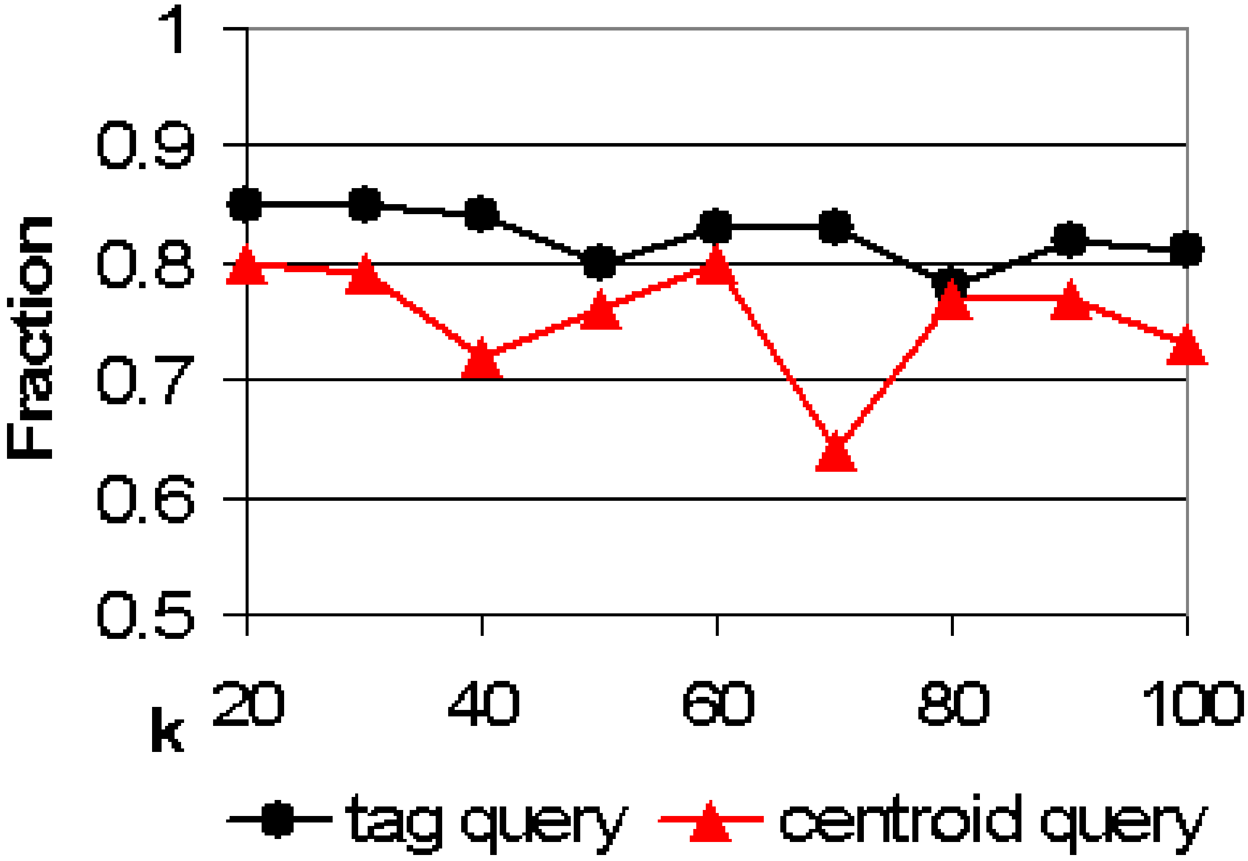
5: For each cluster calculate mean max(sim(a, g)) and mean max(sim(c, g))



query	# queries	# pages
centroid	954	9213
a-tags	883	8633

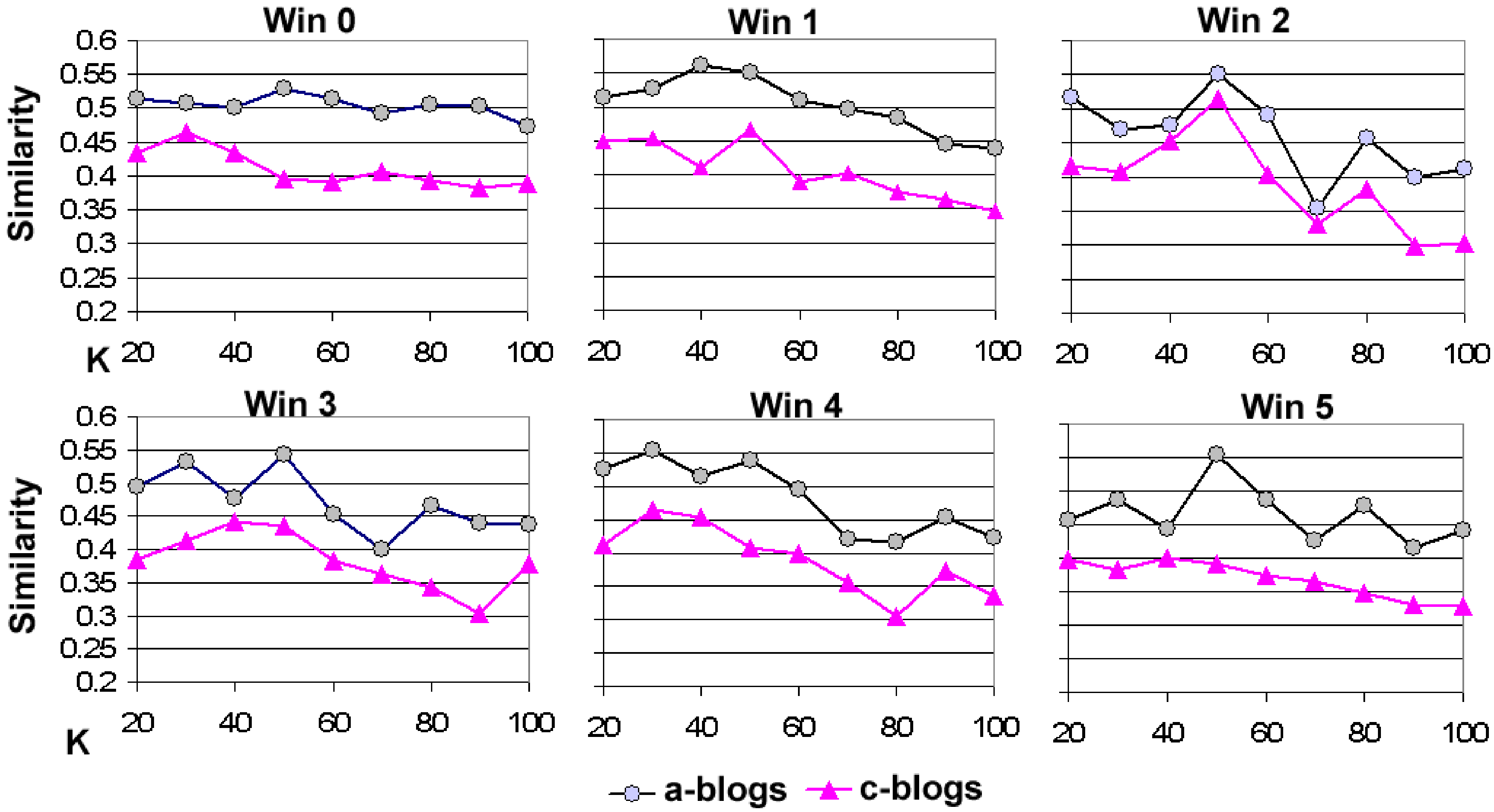
Fraction of clusters

- Where a-blogs are **more similar** than c-blogs than pages retrieved using cluster description



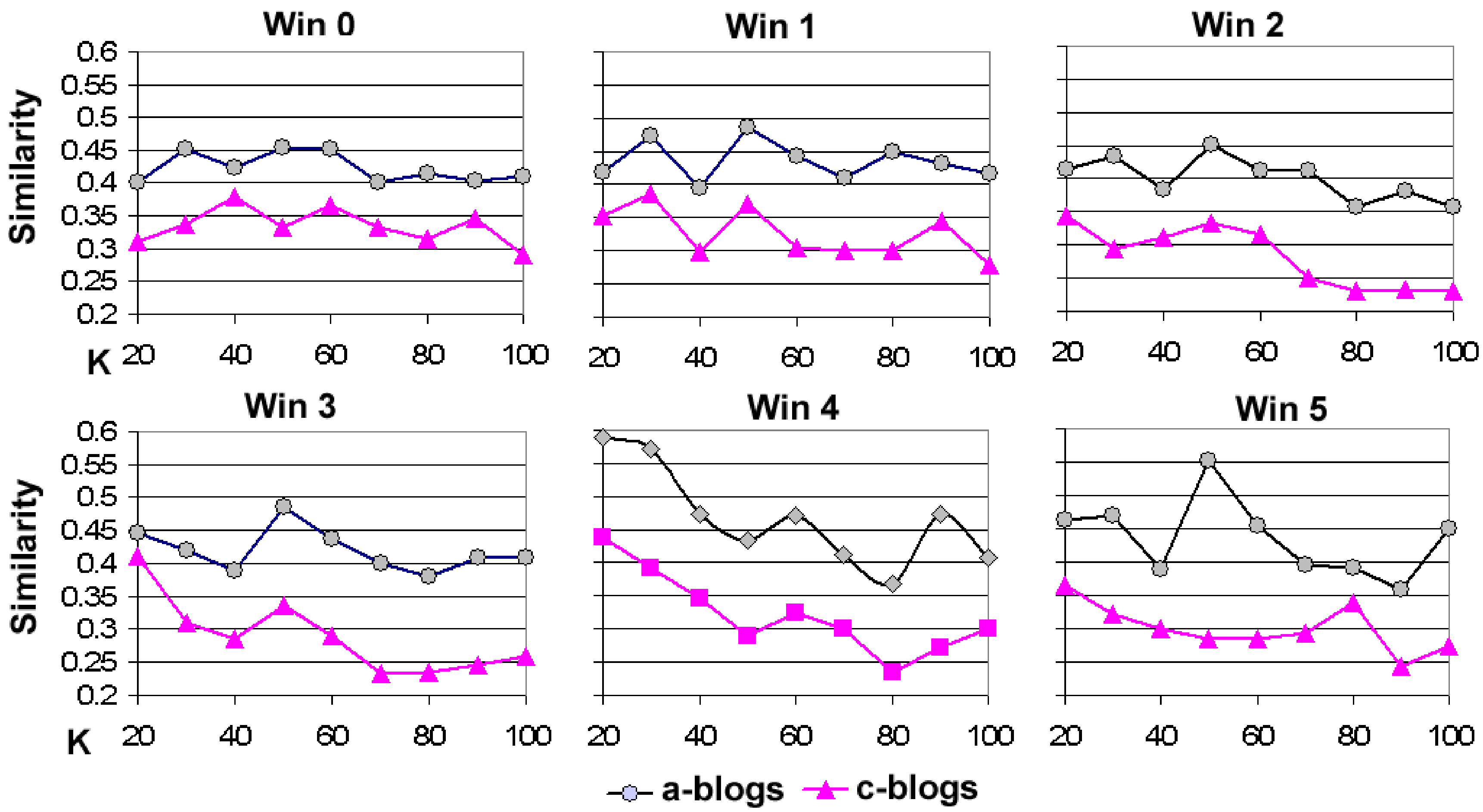
Centroid queries

- Similarity to pages retrieved using cluster **centroid** description



A-tag queries

- Similarity to pages retrieved using cluster **a-tag** description



Results

So, A-tag bloggers tend to

1. Be more similar to each other than c-tag bloggers
2. Be closer to the cluster centroid (topic definition)
3. Be more similar to pages retrieved from Google using the topic description

- Interesting result because a-tag tokens do not contribute to the clustering process
- Yet they consistently allow us to pinpoint those documents that appear to be most relevant to the cluster concept

Example

- Cluster 28 in Win5; k =50
- Cluster description: **mobile, internet, weblog, web, patent**

A-blogs

1) “**Comunications**: technology, economic and social issues at the intersection of telecom, mobility and the Internet”
2) “**IP Blawg**”: *technology and Intellectual property blog*
3) “**Small business IP management blog**: Patent, Trademark, Copyright, Internet, and Technology Law”
4) **“Open Gardens**: Wireless mobility, Digital convergence - Mobile web 2.0”
5) **“Mobile Enterprise Weblog:** the voice of enterprise mobility management”

C-blogs

1) **“Digital Music Den:** Digital Music, online music marketing”
2) **"icarusindie.com – blog about nothing”:** *general computing and technology*
3) **“Dunkie's Saga”** - *personal blog: personal, cars, games, quizzes, some technology*
4) “**Complex Christ** – a vision for church that is organic, networked, decentralized, bottom-up, emergent, communal, flexible, always evolving”
5) **“Philips Brooks patent infringement updates”:** *legal blog on general patent issues (pharmaceutical as well as technological)*

Blogger Drift

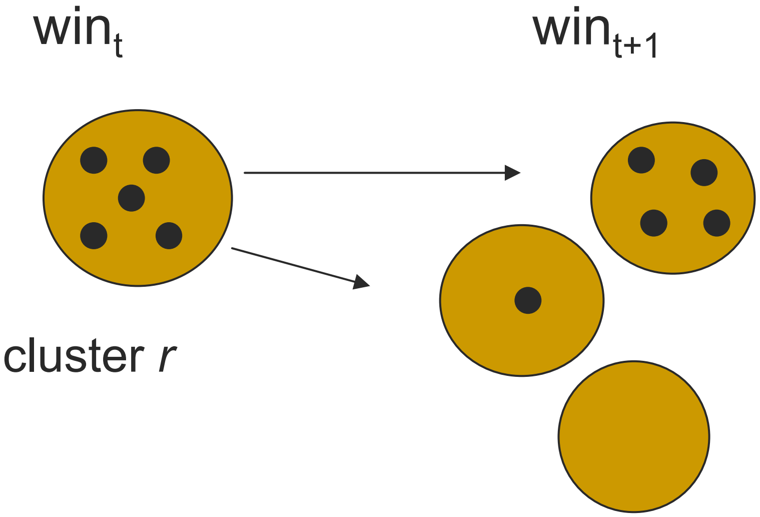
- In previous work we discovered that many bloggers tend to move between topics on a weekly basis
- Problematic : ‘relationships’ established in a clustering will not be useful for very long
- Suggests that many bloggers write in ‘shallow way’ about many topics – even under a single tag
- However, our previous work did not differentiate between a-blogs or c-blogs in each cluster

- The objective of the next experiment is to test how well a-blogs and c-blogs form stable neighbourhoods based on shared topics over time

Blogger entropy

- We define **User Entropy:** a measure of the degree of user dispersion between windows

$$\mathcal{U}_r = -\frac{1}{\log k}\sum_{i=1}^{q}\frac{n_r^i}{n_r}\log\frac{n_r^i}{n_r}$$



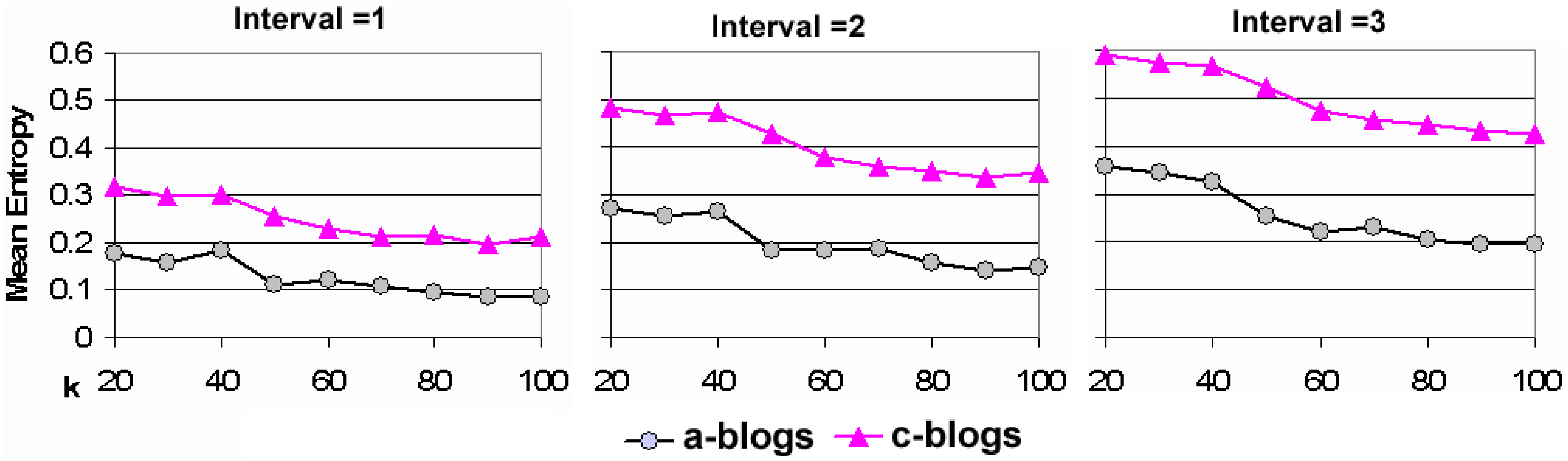
***k* =** number of clusters

q: number of clusters at win_{t+1} containing users from cluster *r*

$\boldsymbol{n_r^i}$: number of users from cluster *r* contained in cluster *i* at win_{t+1}

$\boldsymbol{n_r}$**:** number of users from cluster *r* available at win_{t+1}

Entropy: a-blogs vs c-blogs



- A-blog entropy is lower
- As interval increases a-blogs experience smaller increases in entropy
- Suggests that a-bloggers tend to be write consistently about the same things over time

Conclusions

- We have accumulated empirical evidence to suggest that a-bloggers are topic authorities
  - Tend to form tight subgroups close to cluster topic definition
  - Consistently more similar to pages ranked by Google using the cluster topic definition
  - Tend to stay together at differerent clusterings over time. In other words they tend to write regularly about the same topic
- What characteristics does the a-blogger have?
  - A blogger that is aware of a wider potential readership and chooses his/her tags so that they can be understood easily by others
  - Writes regularly in depth about fairly narrowly defined subjects
  - New professional bloggers

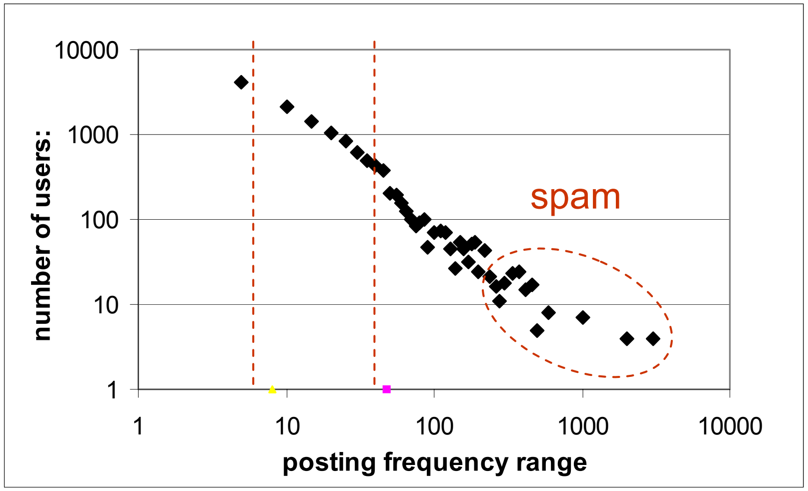
Conclusions 2

- We need to be able to automatically categorize blogs
  - Personal
  - Professional
  - Trend setter
  - ?
- Tag usage is clearly an important feature
- Other features:
  - Topic profiling
  - Linking behaviour: persistent linking to quiz sites
  - Spelling, grammar, language use

Appendix

- 13,518 bloggers: January 16 to February 27, 2006
- Constraints : written in English and tag usage
- Posting frequency follows a power law:
  - 88% of bloggers posted between 1 and 50 times
  - High frequency 'blogs' are generally spam/splog

  - **Data from blogs with posts in range 6-48 : 7549 bloggers (56%)**
  - On average between 1 and 8 posts per week


spam
number of users:
posting frequency range