

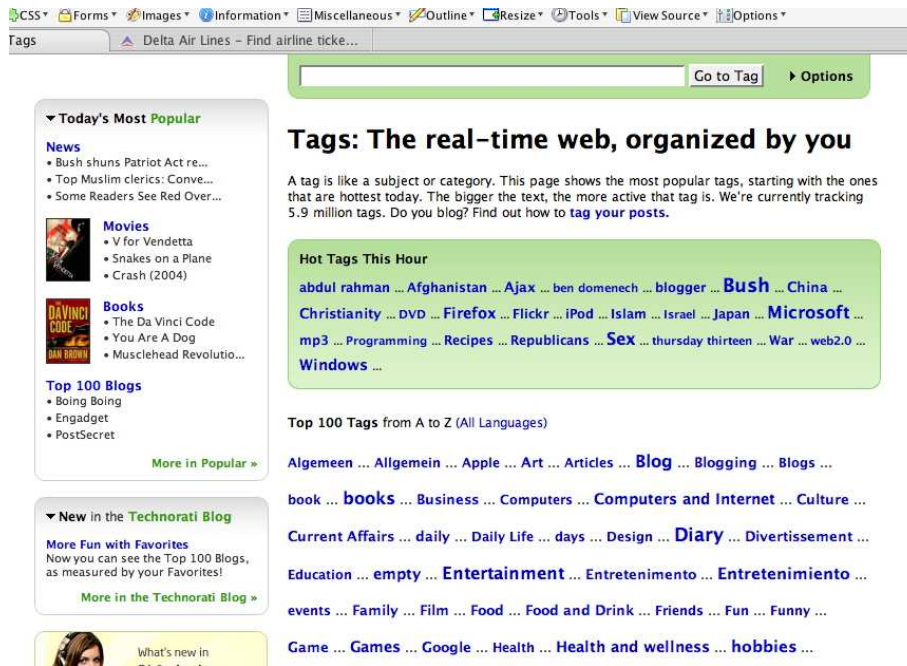
15th International World Wide Web Conference

*Improved Annotation of the Blogosphere via
Autotagging and Hierarchical Clustering*

Chris Brooks and Nancy Montanez

Department of Computer Science
University of San Francisco

Tags



Tags: The real-time web, organized by you

A tag is like a subject or category. This page shows the most popular tags, starting with the ones that are hottest today. The bigger the text, the more active that tag is. We're currently tracking 5.9 million tags. Do you blog? Find out how to [tag your posts](#).

Hot Tags This Hour

abdul rahman ... Afghanistan ... Ajax ... ben domenech ... blogger ... **Bush** ... China ... Christianity ... DVD ... **Firefox** ... Flickr ... iPod ... Islam ... Israel ... Japan ... **Microsoft** ... mp3 ... Programming ... Recipes ... Republicans ... **SEX** ... thursday thirteen ... War ... web2.0 ... Windows ...

Top 100 Tags from A to Z (All Languages)

Algemeen ... Allgemein ... Apple ... Art ... Articles ... **Blog** ... Blogging ... Blogs ... book ... **books** ... Business ... Computers ... Computers and Internet ... Culture ... Current Affairs ... daily ... Daily Life ... days ... Design ... **Diary** ... Divertissement ... Education ... **empty** ... **Entertainment** ... Entretenimento ... Entretenimiento ... events ... Family ... Film ... Food ... Food and Drink ... Friends ... Fun ... Funny ... Game ... Games ... Google ... Health ... Health and wellness ... hobbies ...

- Tagging has recently become a popular method for annotating and organizing blog entries.
- Allows users to attach keywords to blog entries, and share these annotations with others.
- Easy to use and intuitive.
- But what tasks are tags useful for?
- More specifically, do tags help as an information retrieval mechanism?

Shared Tags and Folksonomies

- Tags have (at least) three clear uses:
 - Individual organization
 - Shared annotation of articles into categories
 - Shared annotation as an aid to searching
- We are more interested in tags as a mechanism for sharing information.
- Folksonomy: the meaning associated with a tag will evolve and coalesce through community usage.

Popular tags

About Me, Acne News, Actualite, Actualites, Actualites et politique, Advertising, Allmant, All Posts, amazon, Amigos, amor, Amusement, Anime, Announcements, Articles/News, Asides, Asterisk, audio, Babes, Babes On Flickr, Baby, Baseball, Blogging, Blogs, book, books, Business, Car, Car Insurance, Cars, category, Cell Phones, China, **Cinema**, **Cine cinema**, Comics, **Computadores e a Internet**, **Computer**, **Computers**, **Computers and Internet**, **Computers en internet**, **Computing**, CSS, Curiosidades, Current events, **Data Recovery**, days, Development, diario, Directory, Divertissement, Dogs, dreams, Entertainment, Entretenimento, Entretenimiento, Environment, **etc**, Europe, Event, Everyday, **Everything**, F1, fAcTs, Family, fashion, Feeling, Feelings, FF11, FFXI, **Film**, Firefox, Flash, Flickr, Flutes, Food and Drink, Football, foreign-exchange, Foreign Exchange, Fotos, Friends, Fun, Funny, general, **Game**, **Games**, **Gaming**, Generale, General news, General Posting, General webmaster threads, Geral, Golf, Google, gossip, **Hardware**, Health and wellness, Health Insurance, History, hobbies, Hobby, Home, Humor, Hurricane Katrina, Info, **Informatica e Internet**, International, **Internet**, In The News, Intrattenimento, Java, jeux, Jewelry, jogos, Journal, Journalism, **Juegos**, kat-tun, Katrina, Knitting, Law, Legislation, libros, Life, Links, Live, Livres, Livros, London, Love, Love Poems, Lyrics, **Musica**, Macintosh, Marketing, MassCops Recent Topics, Me, Media, meme, memes, memo, metblogs, metablogging, Military, **Misc**, **Misc.**, **miscellaneous**, MobLog, Mood, **Movie**, **Movies**, murmur, **Music**, **Musica**, **Musik**, **Musings**, **Musique**, **Muziek**, My blog, Nature, **News and politics**, **Noticias e politica**, Opinion, Ordinateurs et Internet, Organizacoes, Organizaciones, Organizaciones, others, **Pasatiempos**, **Passatempos**, PC, Pensamentos, Pensamientos, People, Personal, Philosophy, photo, Pictures, Podcast, Poem, poemas, Poesia, Poker, police headlines, Politik, Projects, Quotes, Radio, Ramblings, random, Randomness, Random thoughts, Rant, Real Estate, Recipes, reflexiones, reizen, Relationships, Research, Resources, Review, RO, RSS, Saude e bem-estar, Salud y bienestar, Sante et bien-etre, School, Science, Search, Sex, sexy, Shopping, Site news, Society, **software**, Spam, Stories, **stuff**, **Tech News**, **technology**, Television, Terrorism, test, Tips, Tools, Travel, Updates, USA, Viagens, Viajes, Video, Videos, VoIP, Votes, Voyages, War, Weather, Weblog, Website, weight loss, **Whatever**, Windows, Wireless, wordpress, words, Work, World news, Writing

The 250 most popular tags on Technorati, as of October 6, 2005

- Things to notice:
 - Tags tend to be general terms
 - Synonyms and related concepts are repeated
 - Misspellings, and different cases
 - Jargon, slang, spam, and Non-English words
 - Non-useful tags (everything, etc, random, test)

Representational Power

- A tag is a label that is applied to a set of blog entries.
- There is no way to specify relationships between tags
 - Opposite, more general/specific, synonym, etc
- In logical terms, tags are a propositional mechanism.
 - This should set off some alarms amongst the AI people in the audience!
- We see users trying to use tags more expressively
 - e.g. “San Francisco, California”
 - This can’t be decomposed, or related to the tag “San Francisco” or the tag “California”
 - Maybe tags are not quite so easy to use ...

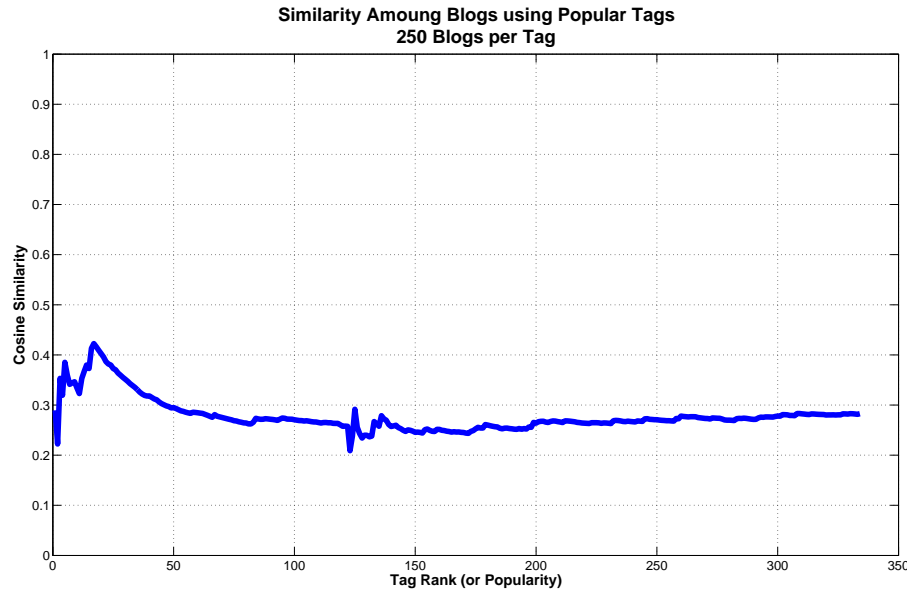
Tags as an Information Retrieval Mechanism

- In this paper, we tried to determine whether tags were useful as an information retrieval mechanism.
 - Specifically, can tags help with a search task?
- How similar are articles that are assigned the same tags?
 - Hypothesis: Rarer tags are better at describing articles than more specific tags.

Tags as an Information Retrieval Mechanism

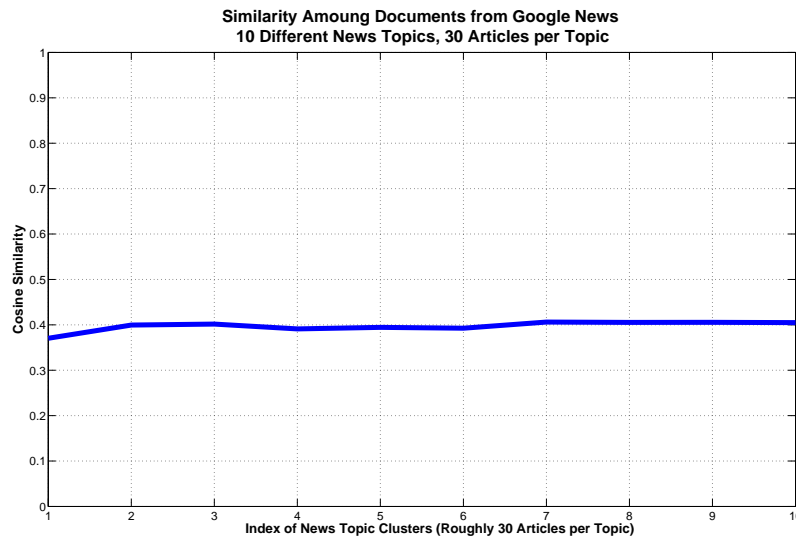
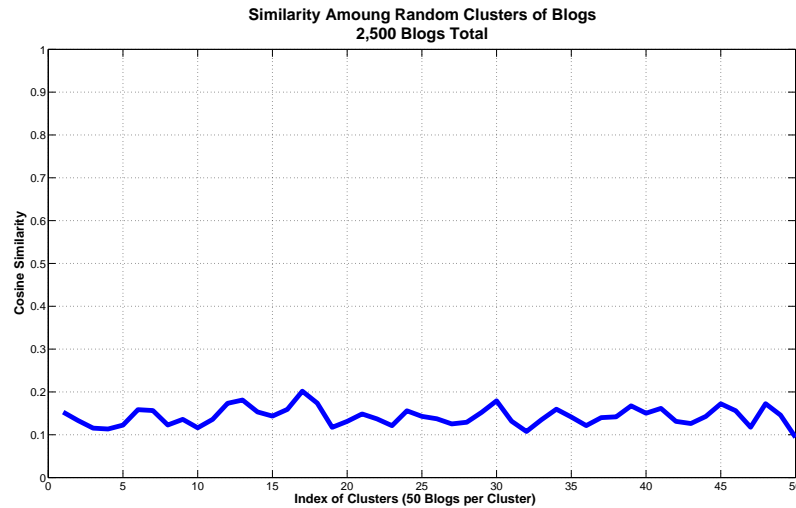
- Retrieved the top 350 tags from Technorati, and then the 250 most recent articles for each tag.
- Articles are converted into weighted vectors, using TFIDF to assign weights to each word.
- All articles that share a tag are assigned to a *tag cluster*
- The size of a tag cluster is measured using the average pairwise cosine similarity.
 - Note: the actual content of the documents is what is evaluated.

How Similar are Tag Clusters?



- Articles with the same tag are somewhat similar.
- Small spike amongst highly popular tags. (game, games, vote)
- Contrary to expectations, articles with rare tags are not more similar than articles with common tags.
- But how similar are these clusters?

Baselines

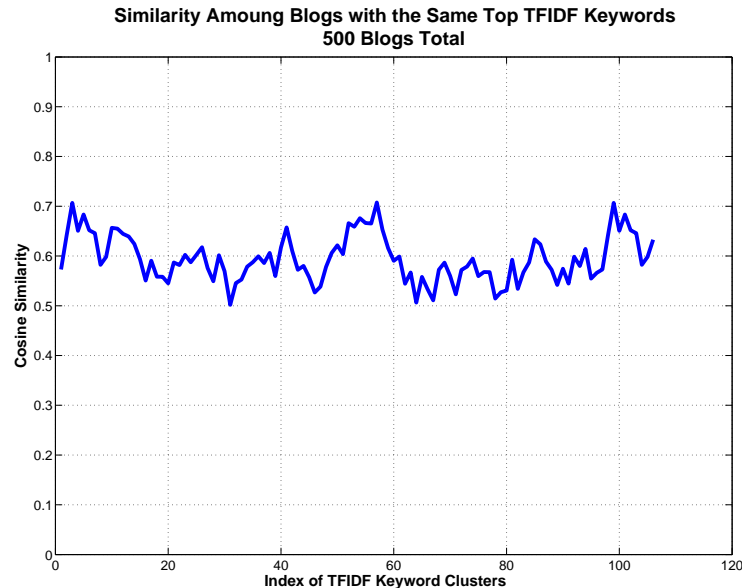


- Tagging clusters articles better than random selection, but worse than Google News.
- Tagging seems most effective at grouping articles into broad topical bins.
- Not very effective as a mechanism for locating particular articles.

Autotagging

- Perhaps users are not very good at choosing tags for search - can automated methods do better?
- Autotagging is the process of automatically assigning tags based on the content of an article.
- Hypothesis: To determine what an article is about, look at the article itself!
- Assign TFIDF scores to all words and extract the highest-scoring words.

Autotagging



Pairwise similarity of clusters of articles sharing a highly-scored word.

- We also extracted the top three highest-scoring words from each article and assigned them as tags.
- Clusters formed using these words were smaller and much more similar than clusters using user-chosen keywords.
- Tags extracted from user text are more helpful in creating specific categories than user-selected tags are.

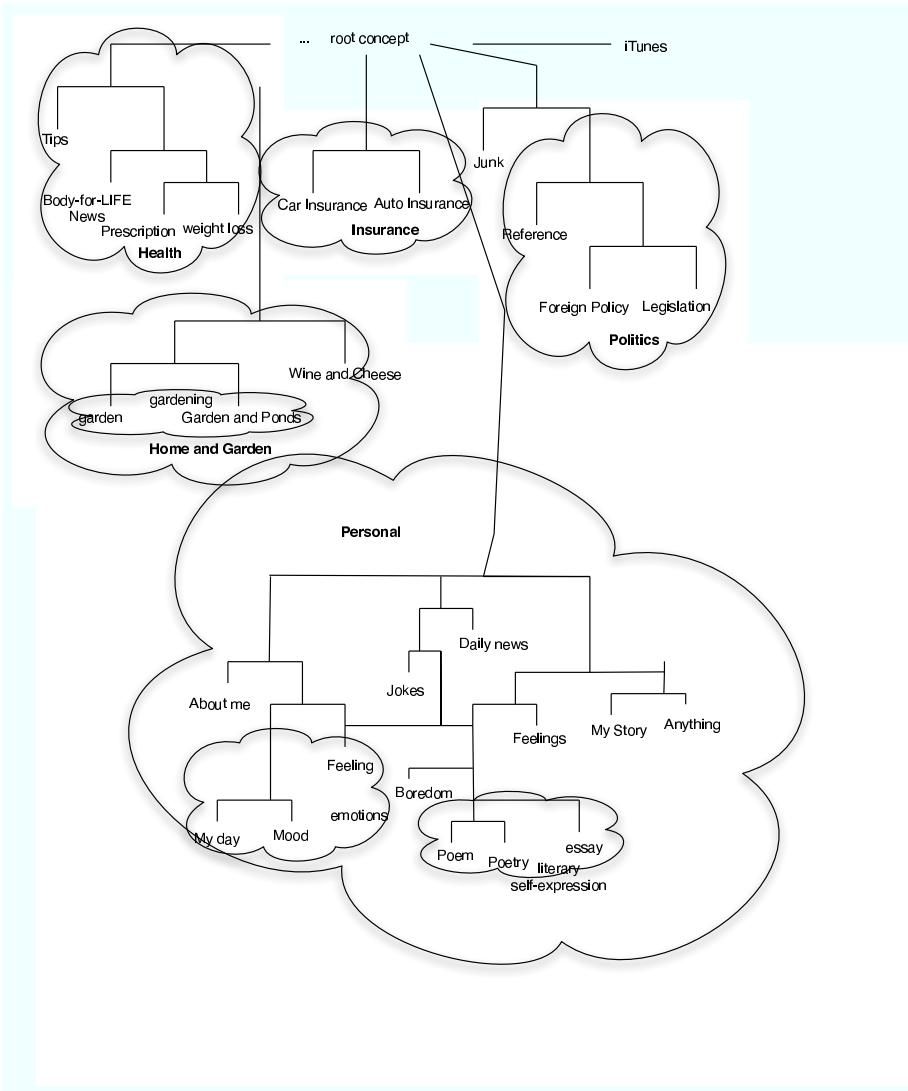
Generating Hierarchies of Tags

- Tags are unable to express related concepts.
- Do related articles have tags judged as similar by a human?
- To address this question, we use agglomerative clustering to construct a tag hierarchy.
- Goal: identify and group tags that are similar or related.

Agglomerative Clustering

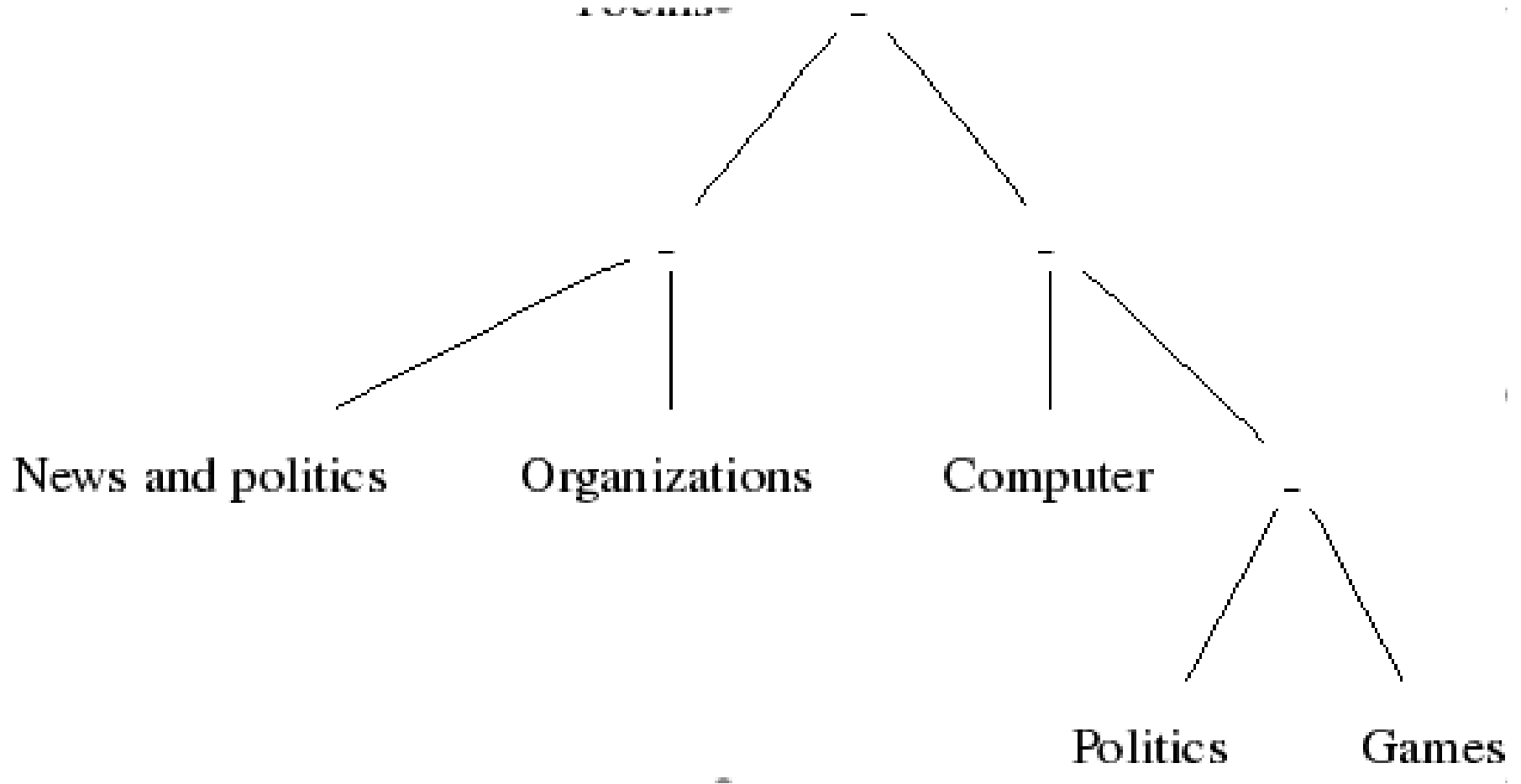
- The agglomerative clustering algorithm is very straightforward:
- Find the two closest tag clusters and merge them into a single abstract cluster. Repeat until one cluster containing all tags remains.
- This yields a dendrogram showing tag similarities.
- $Tags = \{t_1, t_2, \dots, t_n\}$
- **while** $|Tags| > 1$:
 - **find** t_i, t_j **s.t.** $sim(t_i, t_j) \geq sim(t_i, t_k) \forall k \neq i, k \neq j$
 - $t_{new} = t_i \cup t_j$
 - $tags = tags - \{t_i, t_j\}$
 - $tags = tags \cup t_{new}$

Positive results: locating related tags



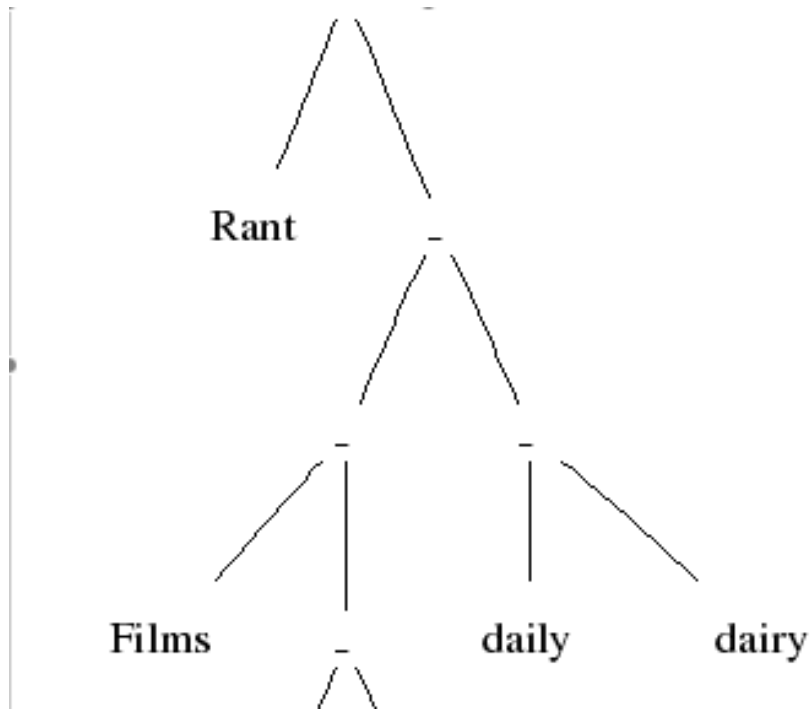
- Clustering is able to construct groups of tags that might be characterized as “related” by a human.

Negative results: shared vocabulary



- Using vectors of single words to represent documents can produce anomalies
 - Both politics and games talk about scores, opponents, and winning.

Negative results: syntactic problems



- “diary” and “dairy” are seen as closely related.
 - Misspelling in the tag.
 - Illustrates a problem with the current representational power of tags.
 - Your tags are only as good as your users!
 - (aside: many community blogs have frequent discussions about “appropriate” tagging vocabularies)

Conclusions

- Tags are very attractive due to their simplicity and ease of use.
- Limited representational power makes them most useful for grouping into large categories.
- By themselves, tags do not seem very effective as a search mechanism.
- Tags can be grouped using clustering techniques, which indicates that relationships can be induced automatically.
- Needed: tools for increasing expressivity without sacrificing ease of use.
 - Expressing relationships, suggesting appropriate tags, catching misspellings, automatically grouping tags.

Future Work

- Current experiments only provide an approximate picture of cluster similarity.
 - Phrase extraction would produce more precise results.
 - Other metrics should be evaluated.
- Currently developing tools that suggest tags based on article similarity and hierarchy.
- Question: do authors and readers use the same tag vocabulary?
- Thanks to Technorati for the use of their data.