

A k-Nearest-Neighbour Method for  
**Classifying** Web Search Results  
with Data in Folksonomies

by Ching-man Au Yeung, Nicholas Gibbins, Nigel Shadbolt

Intelligence, Agents, Multimedia Group  
School of Electronics and Computer Science  
University of Southampton

## The Problem of Ambiguity

- Queries by ambiguous terms return many irrelevant results
- Example: *bridge*
  - 1) a kind of card games;
  - 2) a form of architectural structure;
  - 3) a design pattern in software development;
  - 4) a device in computer networking

# Introduction

The screenshot shows the Delicious website interface. At the top, it says "delicious social bookmarking" and "It's Free! Join Now Sign In". Below this is a blue banner with the text "Your bookmarks will organize themselves. Tag your bookmarks. Collections will naturally emerge." and a "Learn More" link. A search bar is present with the text "Search the biggest collection of bookmarks in the universe...". Below the search bar, there are sections for "Popular Bookmarks" and "Explore Tags". The "Popular Bookmarks" section lists several items with their titles, tags, and the number of saves. For example, "Where to Buy Used Canon Lenses" has 160 saves and tags like "photography", "canon", "used", "shopping", "lenses". Other items include "29 Great Free Textures | Abduzeedo - design inspiration & tutorials" (184 saves), "JeffBridges.com - Ironman book" (149 saves), "Essential free apps for your web design toolkit | News | TechRadar UK" (214 saves), "Toxel.com » 24 Beautiful and Creative Website Headers" (228 saves), "Seadragon Ajax - Microsoft Live Labs" (192 saves), "MAKE: Blog: Arduino Gift Guide!" (160 saves), "27 Free Must-have Online Collaboration Tools - Crazeegeeckchick.com" (328 saves), and "is it going to rain?" (395 saves). A "Popular Tags" list is also visible on the right side.

Delicious

The screenshot shows the BibSonomy website interface. At the top, it says "BibSonomy :: search:all ::" and "A free social bookmark and publication sharing system." Below this is a navigation bar with links for "tags", "relations", "groups", "popular", "username:", and "password:". There is also a "help · blog · about" and "login · register" section. The main content area features a paragraph about BibSonomy: "BibSonomy is a system for sharing bookmarks and lists of literature. When discovering a bookmark or a publication on the web, you can store it on our server. You can add tags to your post to retrieve it more easily. This is very similar to the bookmarks/favorites that you store within your browser. The advantage of BibSonomy is that you can access your data from wherever you are. Furthermore, you can discover more bookmarks and publications from your friends and other people." Below this is a section for "bookmarks" and "publications" with filters like "RSS", "HTML", "PDF", "more". The "bookmarks" section lists items such as "Définitions dans le cadre des soins palliatifs", "Search Engine Optimization and Keywords", "Yahoo: 'Searches more sophisticated and specific' | Digital Markets | ZDNet.com", "Women Kick Guy In The Nuts", "Maus Spiele - Die Sendung mit der Maus - WDR Fernsehen", and "Secure Moodle - liitmwiki". The "publications" section lists items such as "Foundations of Therapeutic Interviewing", "XML: kurz und gut", "Learning Graph Walk Based Similarity Measures for Parsed Text", "Beginning ASP.NET 3.5 in VB 2008: from novice to professional", "The cost of a 60% cut in CO2 emissions from homes: what do experience curves tell us?", and "Does it matter who contributes: a study on featured articles in the german wikipedia". A "filter" section on the right lists various tags like "alpha", "freq", "cloud", "list", "2008", "ajax", "analysis", "api", "application", "architecture", "archive", "art", "article", "authentication", "awareness", "barberg", "bibliography", "bibsonomy", "bibtex", "blog", "book", "code", "collaboration", "communication", "community", "concept", "conference", "contest", "crawl", "data", "database", "design", "deutschland", "development", "digital", "download", "edge", "economics", "editor", "folksonomy", "free", "history", "howto", "information", "internet", "java", "language", "linux", "maps", "marketing", "math", "mathematics", "metadata", "model", "music", "network", "networks", "nettags", "nlp", "of", "ontology", "opensource", "optimization", "php", "programming", "rdf", "repository", "research", "retrieval", "science", "search", "security", "semantic", "semanticweb", "service", "social", "evaluation".

BibSonomy

## Collaborative Tagging Systems

- ◆ Aggregate user-contributed metadata of Web resources
- ◆ Provide rich information about the relations between different tags
- ◆ Sources for understanding how keywords are used on the Web

## Multiple Meanings of Tags

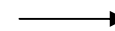
- ◆ Tags have multiple meanings, or they are used in different contexts
- ◆ It is possible to extract related tags in different contexts
- ◆ E.g. sf:  
  
(california, bayarea, travel, ...) → San Francisco  
(scifi, fantasy, fiction, ...) → Science Fiction

## Our Proposal

- ◆ Building classifiers using data in folksonomies:

### Wikipedia page of San Francisco

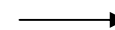
The City and County of San Francisco is the fourth most populous city in **California** and the 14th most populous city in the United States ... Among the most densely populated cities in the country, San Francisco is part of the San Francisco **Bay Area** ... The city is located at the tip of the San Francisco Peninsula, with the Pacific Ocean to the west, ...



San  
Francisco

### Wikipedia page of Science Fiction

Science fiction (abbreviated SF or **sci-fi** with varying punctuation and capitalization) is a broad genre of **fiction** that often involves speculations based on current or future science or technology. Science fiction is found in books, art, television, films, games, theatre, and other media ... this includes **fantasy**, horror, and related genres.

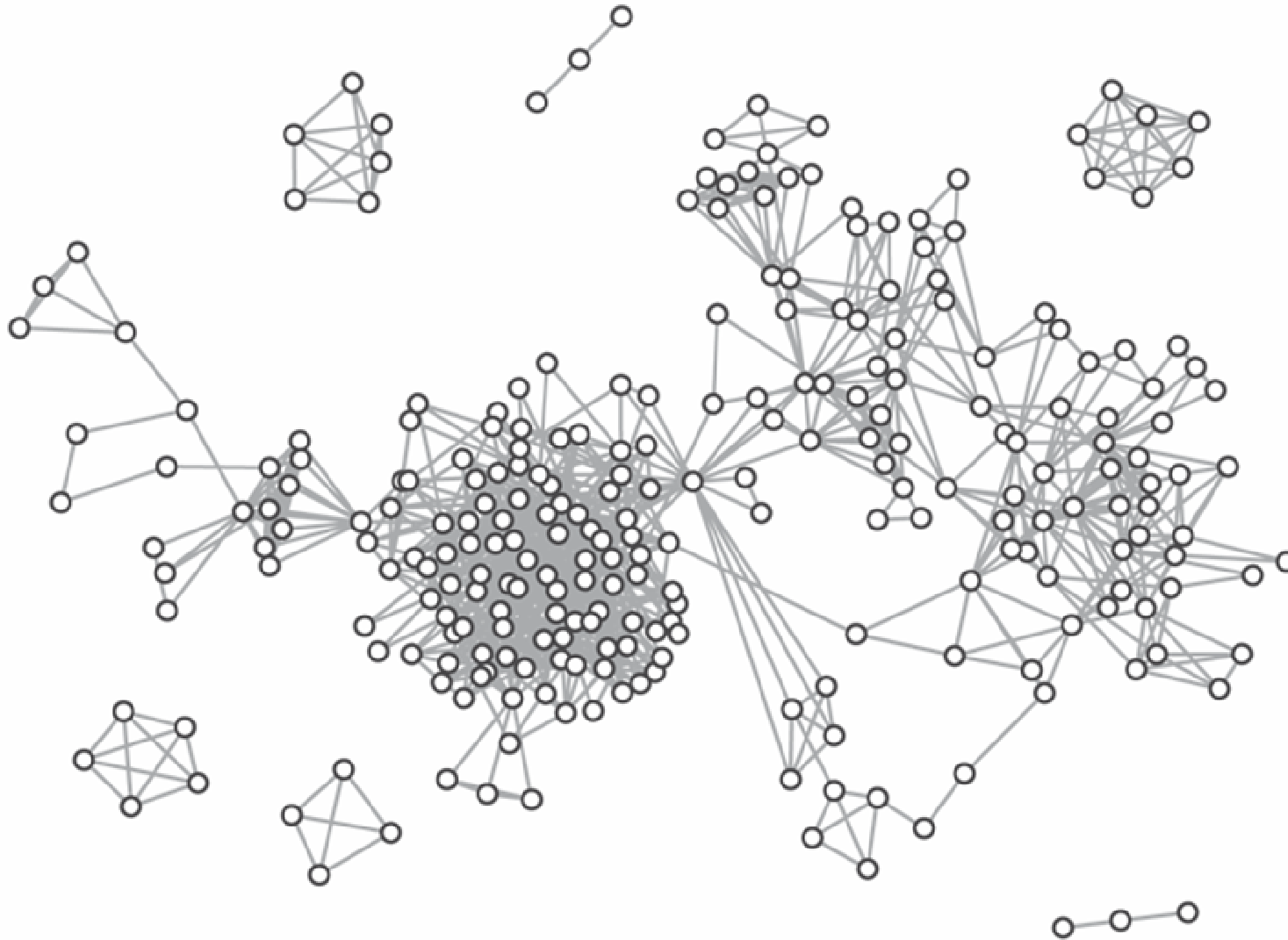


Science  
Fiction

## Clustering of Folksonomy Networks

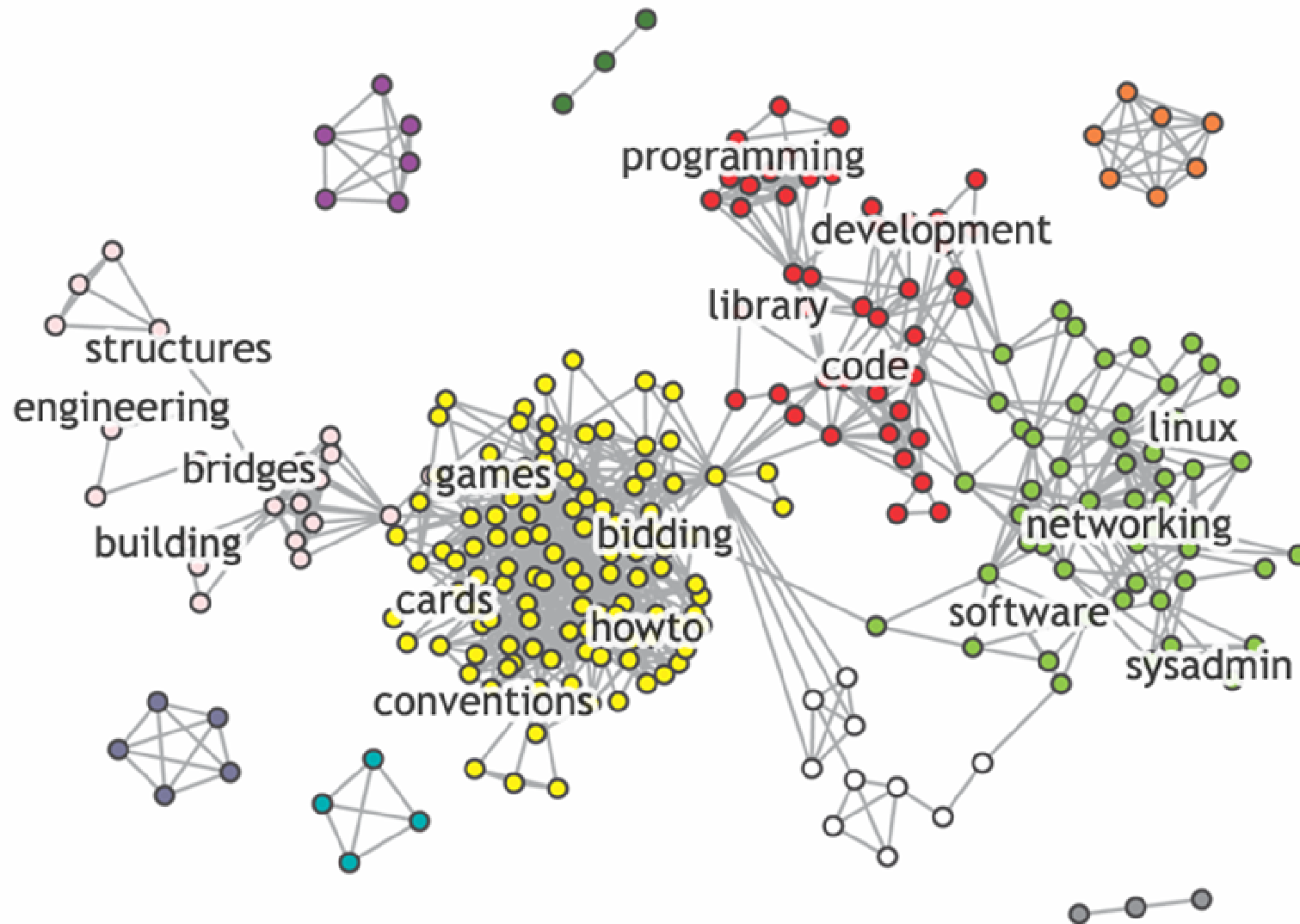
- ◆ Construct a document network from a folksonomy
- ◆ Cluster documents based on the users who have used the tag on the documents  
(the community-discovery algorithm described in [Newman 2004] is used in this paper)
- ◆ Extract frequently co-occurred tags as representations of the different classes (meanings)

# Clustering of Folksonomy Networks





# Clustering of Folksonomy Network



# Clustering of Folksonomy Network

**Design pattern** bridge, programming, development, library, code, ruby, tools, software, adobe, dev

---

**Card game** bridge, games, cards, game, imported, howto, conventions, card, bidding, online

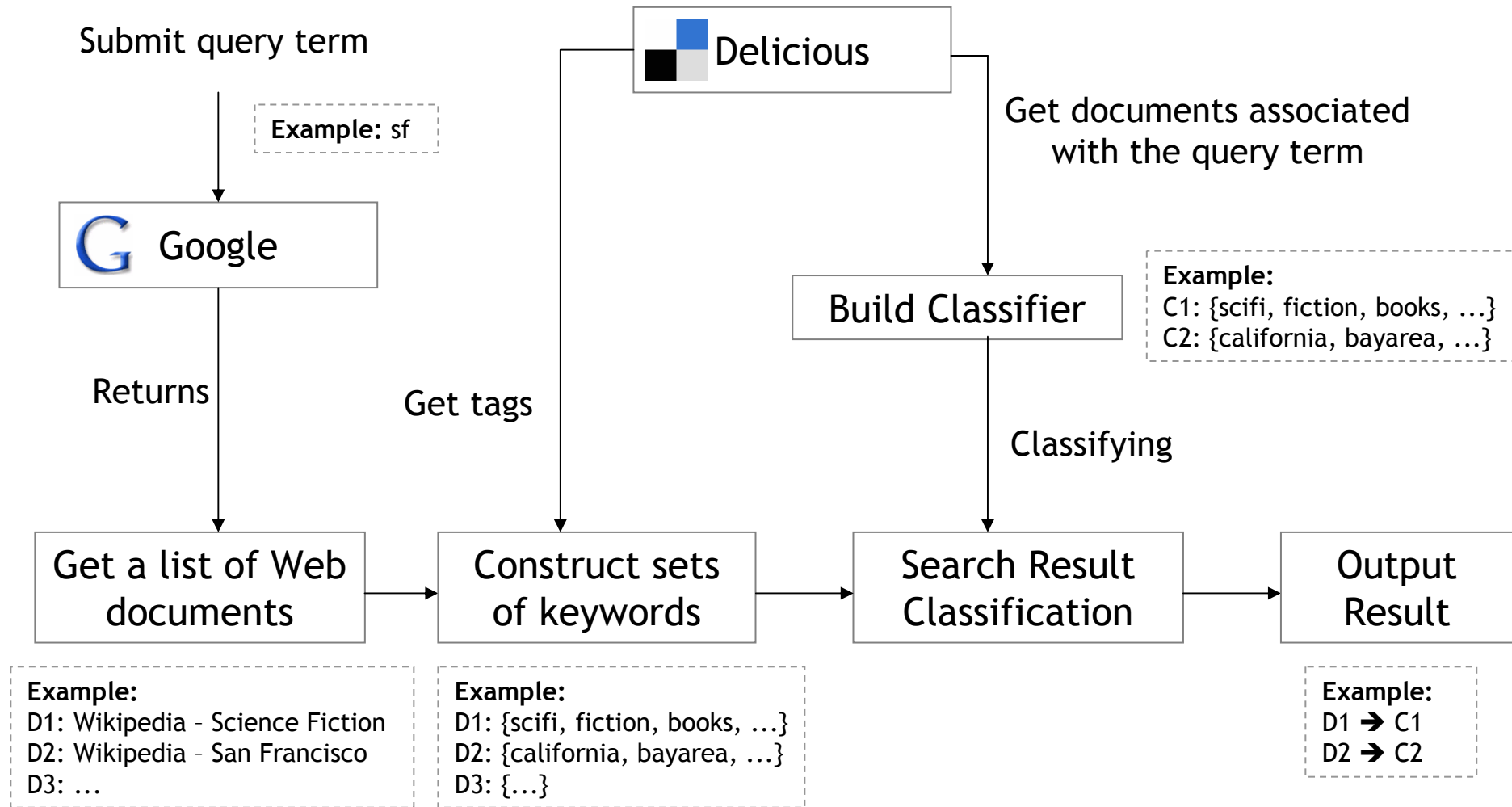
---

**Computer networking** bridge, networking, linux, network, howto, software, sysadmin, firewall, virtualization, security

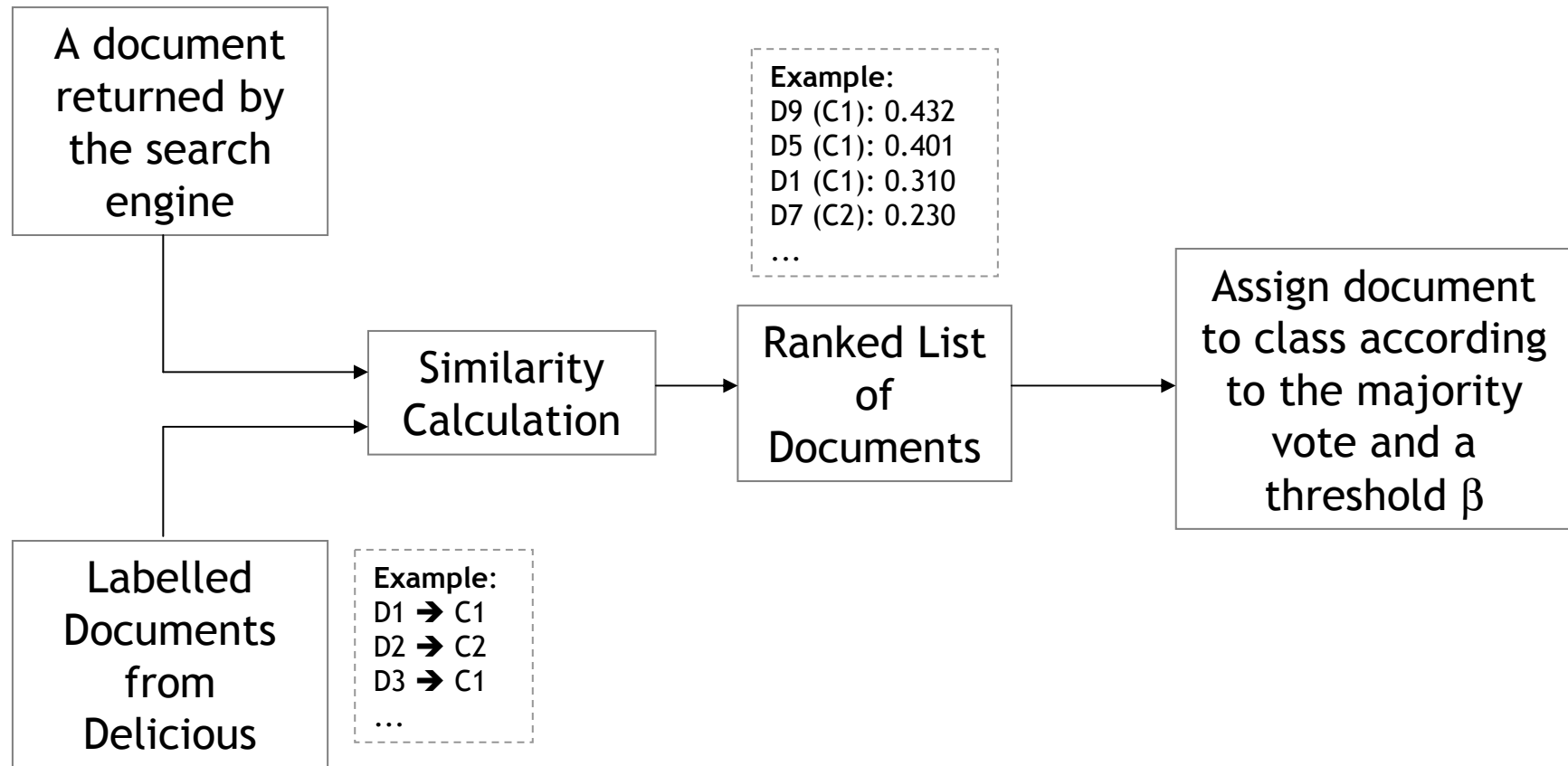
---

**Architecture** bridge, bridges, structures, engineering, science, physics, school, education, building, reference

# Web Search Result Classification



## k-Nearest-Neighbour Classifier



## Data Preparation

- ◆ Ten tags which are used in multiple contexts in Delicious are chosen
- ◆ Documents associated with the tags as well as the users who assigned the tags are retrieved
- ◆ Testing dataset obtained by submitting query to Google and obtaining the top 50 pages returned

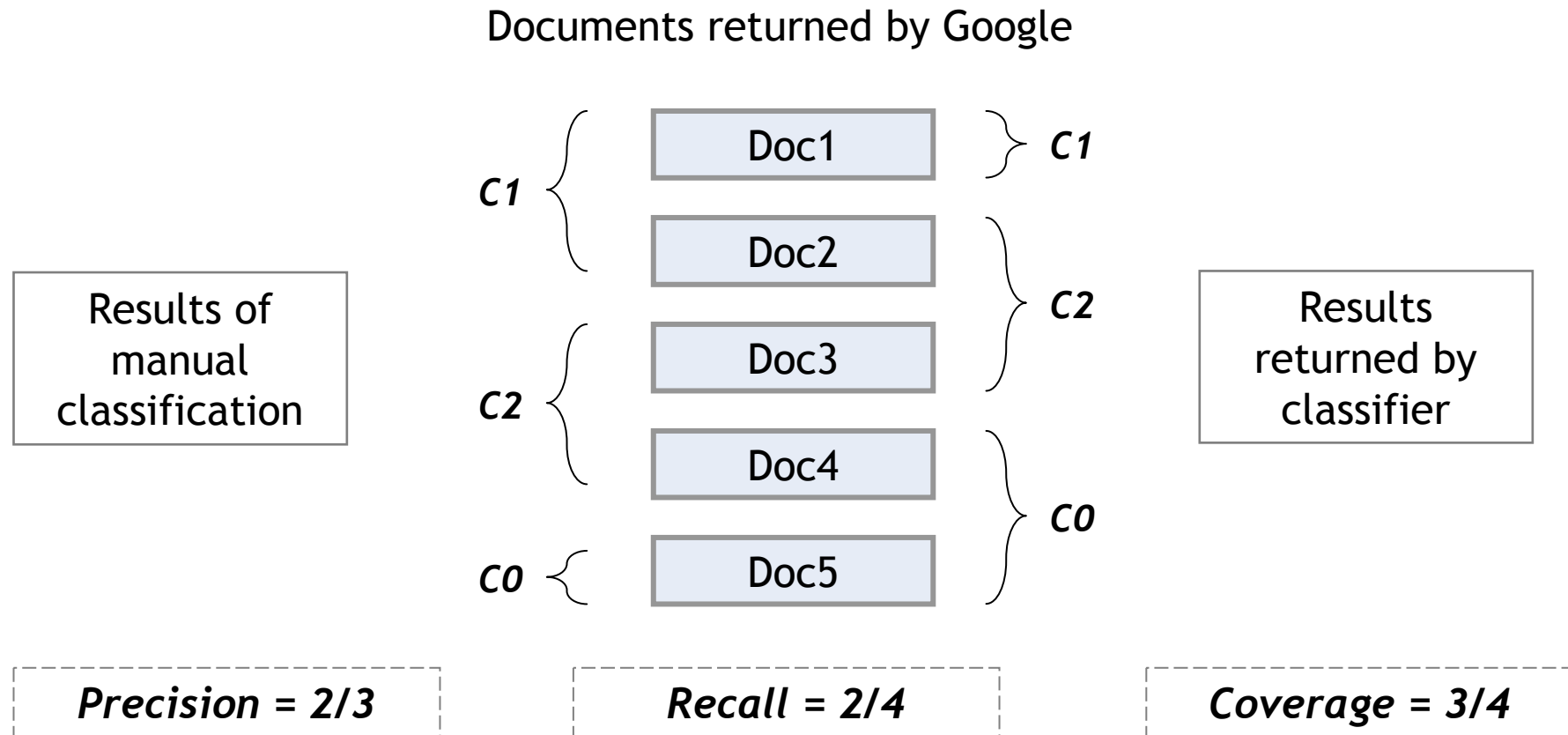
## Classifiers Built

Tag	Context	Class Label
sf	San Francisco	sanfrancisco, bayarea, san, francisco, california, travel, events
	Science fiction	scifi, fiction, books, sci-fi, literature, writing, science, fantasy
soap	Cleaning agent	soapmaking, diy, recipes, crafts, shopping, making, howto
	Web services	webservices, webservice, programming, web, xml, soa, java
wine	Software application	linux, ubuntu, howto, windows, software, tutorial, emulation
	Beverage	food, shopping, drink, vino, cooking, alcohol, blog, news
xp	Windows XP	windows, software, tools, pc, computer, tech, winxp, microsoft
	Extreme programming	software, programming, process, methodology, development

## Performance Measures

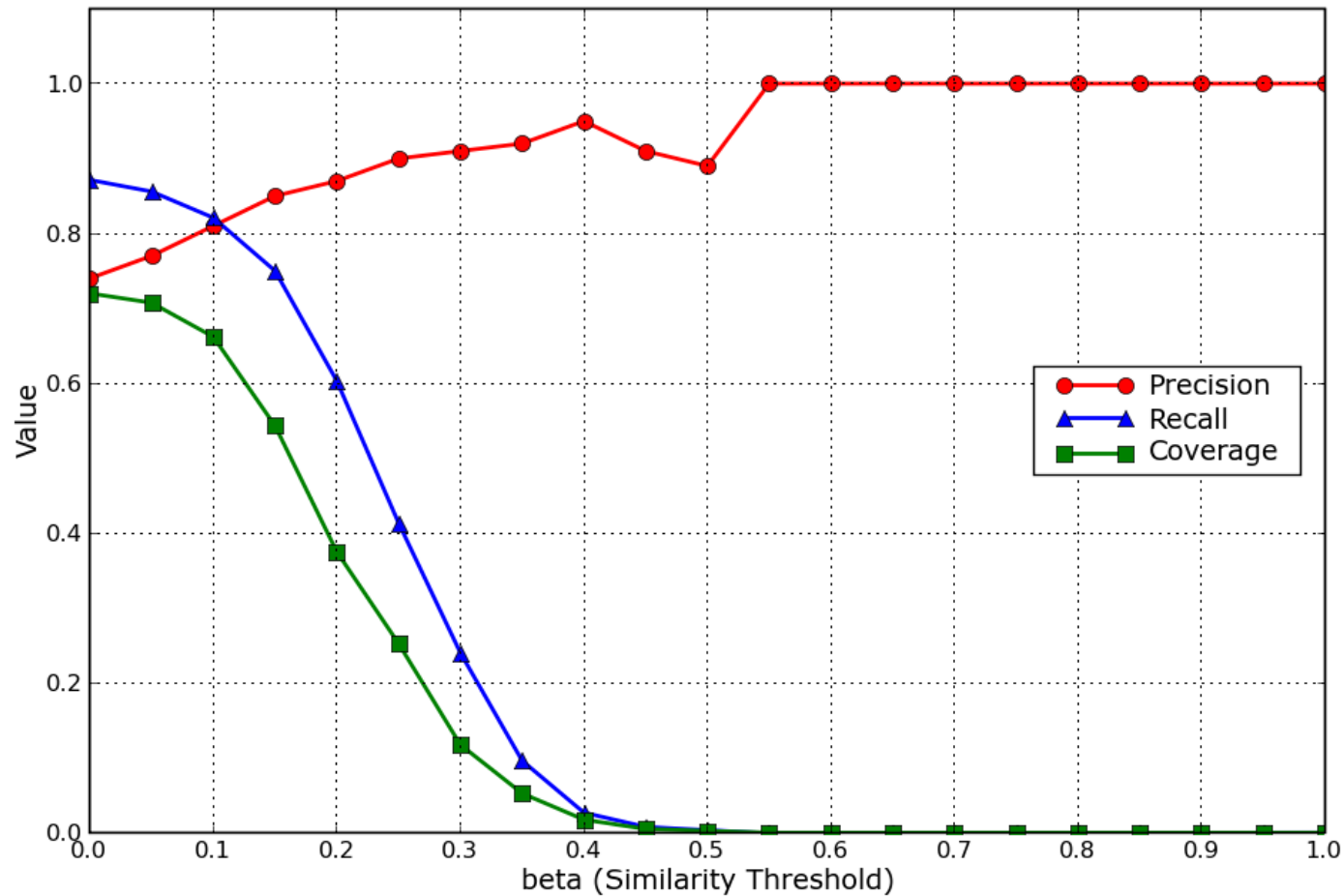
- ◆ ***Precision***  
Measures the percentage of documents which are classified correctly.
- ◆ ***Recall***  
Measures the percentage of classifiable documents which are classified correctly
- ◆ ***Coverage***  
Measures the proportion of documents the classifiers are able to classify

## An Example



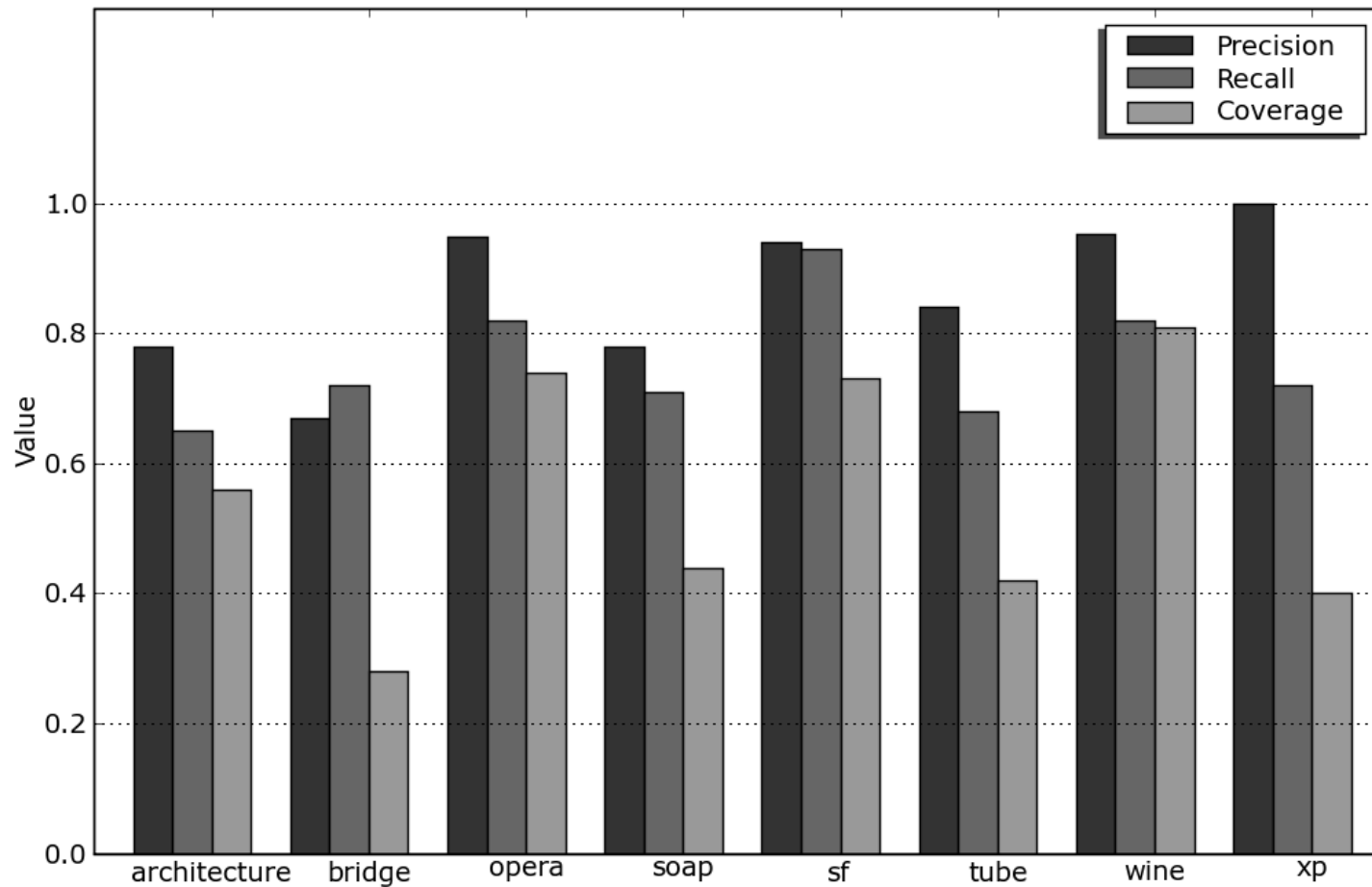


# Evaluation



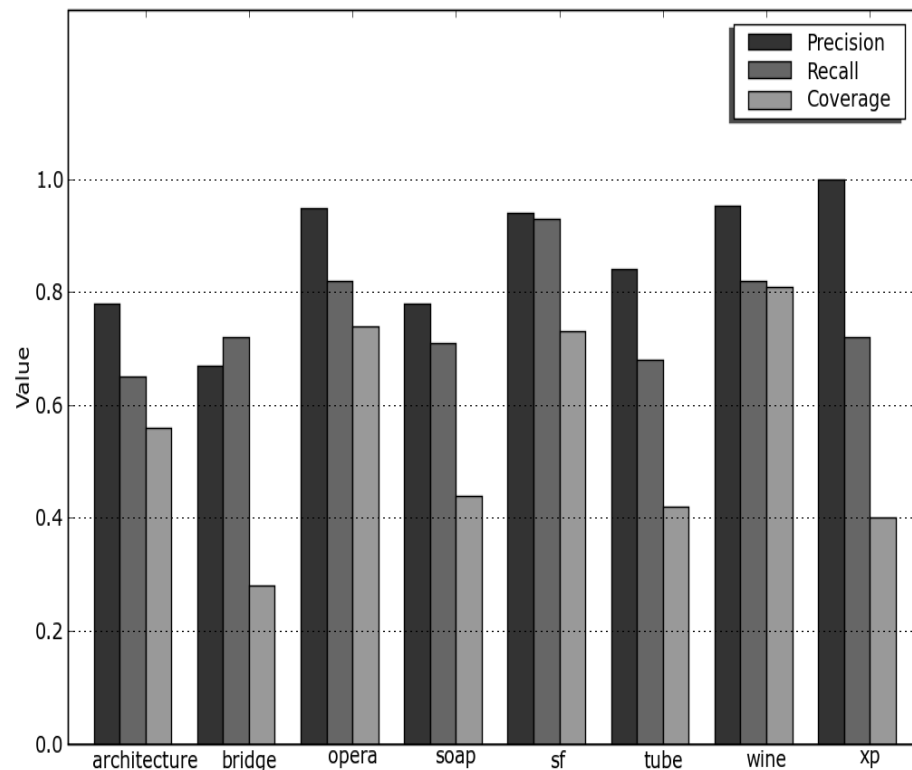
*Figure 1. Precision, recall and coverage against different values of  $\beta$ .*

# Evaluation



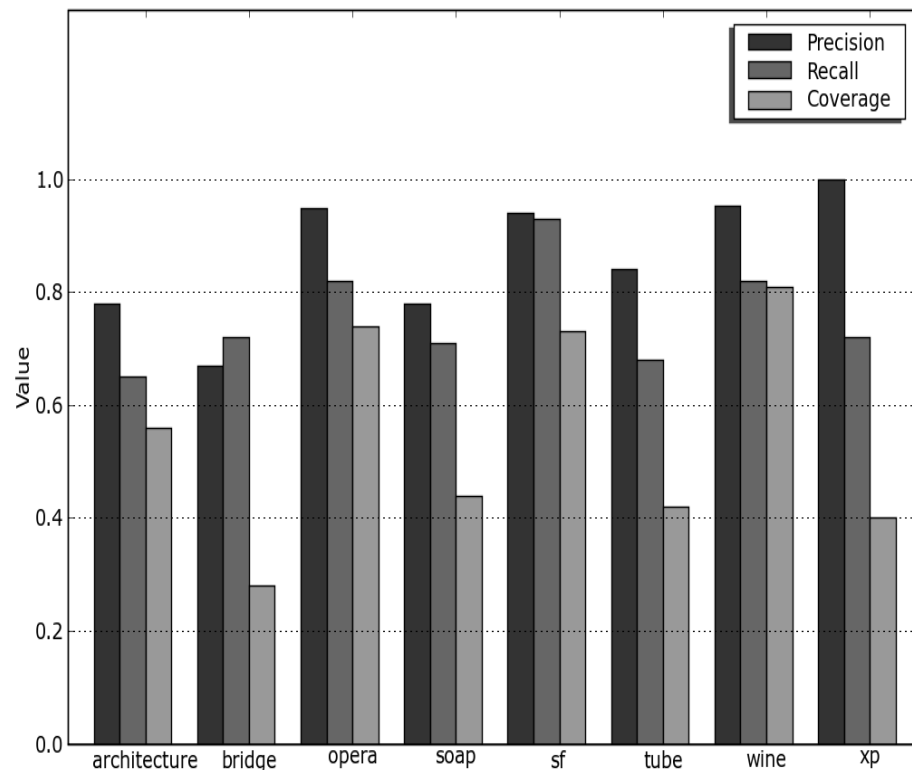
**Figure 2. Precision, recall and coverage for different tags.  
( $k = 11$ ,  $\beta = 0.15$ )**

## Precision



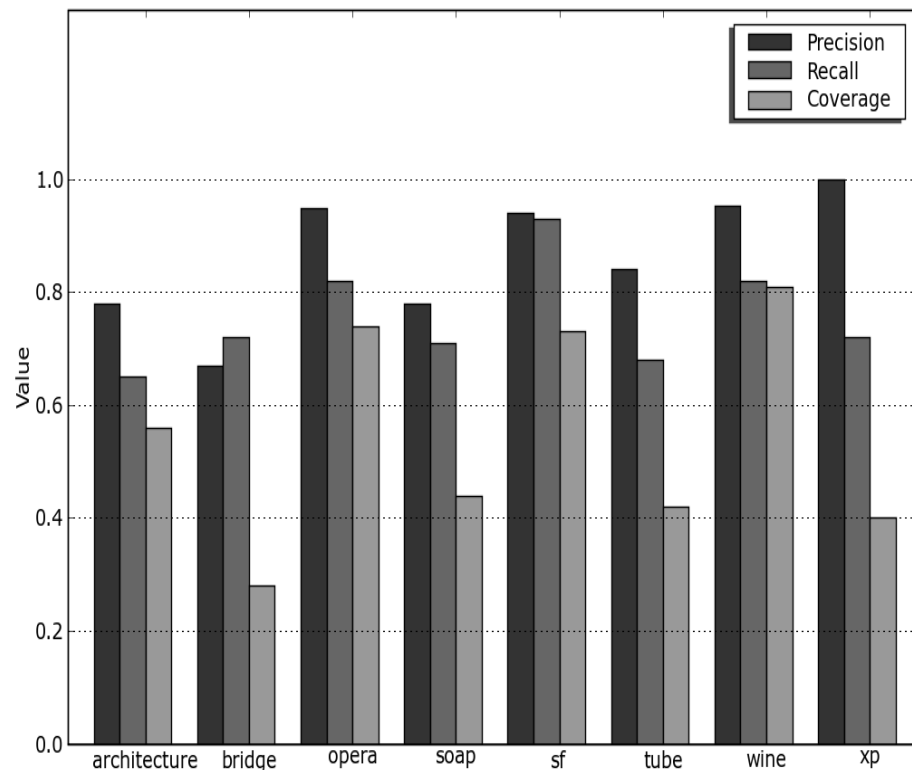
- ◆ Precision is generally quite high (67-100%)
- ◆ Clustering process provides good basis for the kNN classifier
- ◆ Low precision cases: different keywords even for same context

## Recall



- ◆ Recall ranges from 65% to 93%
- ◆ Some documents cannot be classified (recognised)
- ◆ Mainly because that keywords do not match well

## Coverage



- ◆ Has the largest range: 28-81%
- ◆ Due partly to low recall
- ◆ Some contexts not discovered by the clustering process (e.g. tube)
- ◆ Also, there are irrelevant results (e.g. bridge)

# C onclusions

- ◆ Folksonomies offer rich information on the relations and semantics of tags, and can be used to enhance Web search
- ◆ Advantages over using of dictionaries or thesauruses (able to keep up with new meanings)
- ◆ Future research directions:
  1. Building more comprehensive classifiers
  2. Use of other clustering methods
  3. Larger scale of evaluation

**Thank You!**

Albert Au Yeung  
cmay06r@ecs.soton.ac.uk  
<http://users.ecs.soton.ac.uk/cmay06r/>