

# Exploring Social Annotations for the Semantic Web

Xian Wu  
Lei Zhang  
Yong Yu

Shanghai Jiao Tong University  
IBM China Research Lab  
Shanghai Jiao Tong University

# Outline

- ❖ **Background & Motivation**
- ❖ **Exploring Social Annotations**
- ❖ **Semantic Search and Discovery**
- ❖ **Evaluation**

# Semantic Annotations

- ❖ The resources in WWW are required to be annotated in machine-understandable metadata for automatic process.
- ❖ Primary approaches achieves in a top-down manner:
  - Define an ontology first.
  - Use the ontology to add semantic markups to web resources.
  - The markups are usually written in standard languages such as RDF and OWL.
  - The semantics is provided by the ontology which is shared among different web agents and applications.

# Disadvantages of Top-Down Approaches

## ❖ **Negotiation**

It's difficult to establish a common ontology for large scale distributed web resources to satisfy users with all kinds of background.

## ❖ **Evolution**

Even if the consensus of a common ontology is achieved, it's difficult to catch the fast pace of change of web resources and users' vocabulary.

## ❖ **High Barrier**

Using common ontology to annotate web resources requires background skill in ontology engineering, thus it has a high barrier to entry.

# The Emergence of Social Annotation

## ❖ **Social Participating in WWW**

In recent years, normal web users contribute more and more to WWW in the form of blog, social bookmarks and so on.

## ❖ **Annotation Freely**

Normal web users can also take part in to annotate web resources with the help of social annotation services. They can annotate their bookmarks (Delicious), photos (Flickr), wishes (43things)...freely with any keywords they like and share them with other users.

# Social Annotations

## ❖ Advantage

- No common ontology or dictionary are needed
- Easy to access
- Sensitive to information drift

## ❖ Disadvantage

- Ambiguity Problem: For example, “XP” can refer to either “Extreme Programming” or “Windows XP”.
- Group Synonymy Problem: two seemingly different annotations may bear the same meaning.

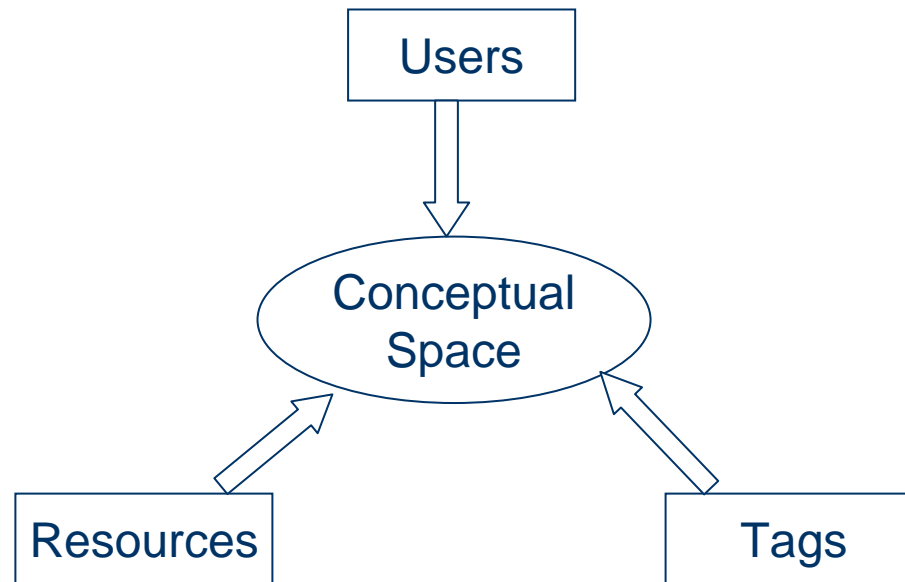
- ❖ **Background & Motivation**
- ❖ **Exploring Social Annotation**
- ❖ **Semantic Search and Discovery**
- ❖ **Evaluation**

# Deriving Emergent Semantics

- ❖ The social annotation data can be abstracted to a set of triples: ***{user, resource, tag}***
  - Users:  $U = \{ u_1, u_2, \dots, u_K \}$
  - Resources:  $R = \{ r_1, r_2, \dots, r_M \}$
  - Tags:  $T = \{ t_1, t_2, \dots, t_N \}$
- ❖ Implicit semantics are embedded in the frequencies of co-occurrences of user, tag and resources.
  - Tags are usually semantic related to each other if they are used to tag the same or related resources for many times or they are used by the same user or users with similar interests for many times. Resources and users are in like manner.
  - The frequencies of co-occurrences give expression to the implicit semantics embedded in them.

# Representation of Semantics

- ❖ We represent semantics of an entity (a web resource, a tag or a user) as a multi-dimensional vector where each dimension represents a special category of knowledge.
- ❖ Every entity can be mapped to a multi-dimensional vector, whose component measures the relativity between the entity and the category of knowledge. If one entity relates to a special category of knowledge, the corresponding dimension of its vector has a high score.
- ❖ The total knowledge of users, tags and resources are the same, we can represent them in the same multi-dimensional space, which we call conceptual space.



# Statistical Co-occurrence Model

- ❖ There are researches on the statistical analysis of co-occurrences of objects.
  - Develop parametric models
  - Estimate parameters by maximizing log-likelihood on the existing data set.
- ❖ We extend the bipartite Separable Mixture Models [Hofmann, 98] to tripartite model, and then use the model to process social annotation data.

# Statistical Co-occurrence Model(2)

- ❖ The generation of social annotation data can be modeled in the following probabilistic process:
  - Choose a dimension  $d_\alpha$  to represent a category of knowledge according to the probability  $p(d_\alpha)$
  - Measure the relativity between the interest of user  $u_i$  and the chosen dimension with the conditional probability  $p(u_i / d_\alpha)$
  - Measure the relativity between the semantics of a resource  $r_j$  and the chosen dimension with conditional probability  $p(r_j / d_\alpha)$
  - Measure the relativity between the semantics of a tag  $t_k$  and the chosen dimension according to the conditional probability  $p(t_k / d_\alpha)$
- ❖ Using the EM methods to maximizing log-likelihood on social annotation data set, the parameters above can be acquired.

# Vector Values

- ❖ With the acquired parameters, the vector value of a tag  $t_k$  can be calculated as:

$$p(d_\alpha | t_k) = \frac{p(t_k | d_\alpha) p(d_\alpha)}{p(t_k)} \propto p(t_k | d_\alpha) p(d_\alpha)$$

- ❖ Since  $\sum_{\alpha=1}^D p(d_\alpha | t_k) = 1$ , we are able to calculate the vector values of tags, the vector values of users and resources can be calculated in the same way.

# Experiment Data

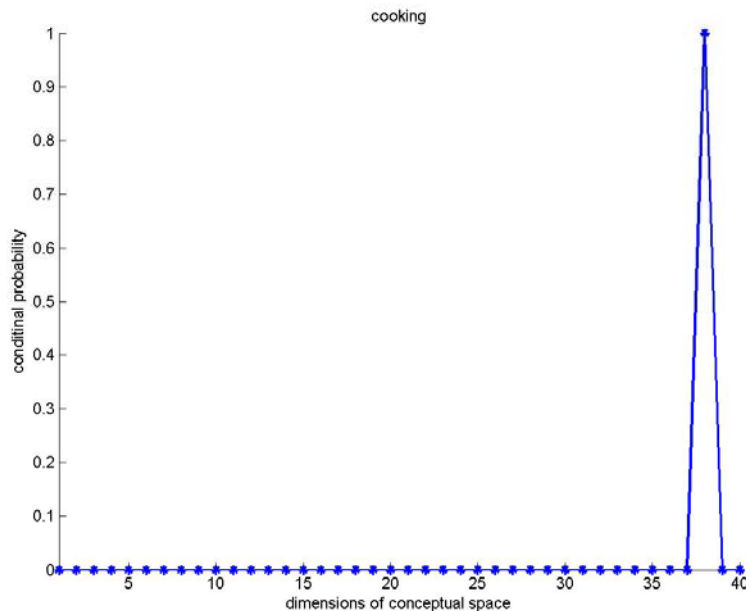
- ❖ We collected a sample of Delicious data by crawling its website during March 2005.
- ❖ Raw Data Set:
  - 2,879,614 triples made by 10,109 users on 690,482 URLs with 126,304 tags.
- ❖ Refined Data Set:
  - 907,491 triples made by 8676 users, 9770 tags and 16011 URLs.
- ❖ In our experiment, we set the number of dimensions in conceptual space to 40 which perform well on the experiment data set.

# Top5 tags in Each Dimension

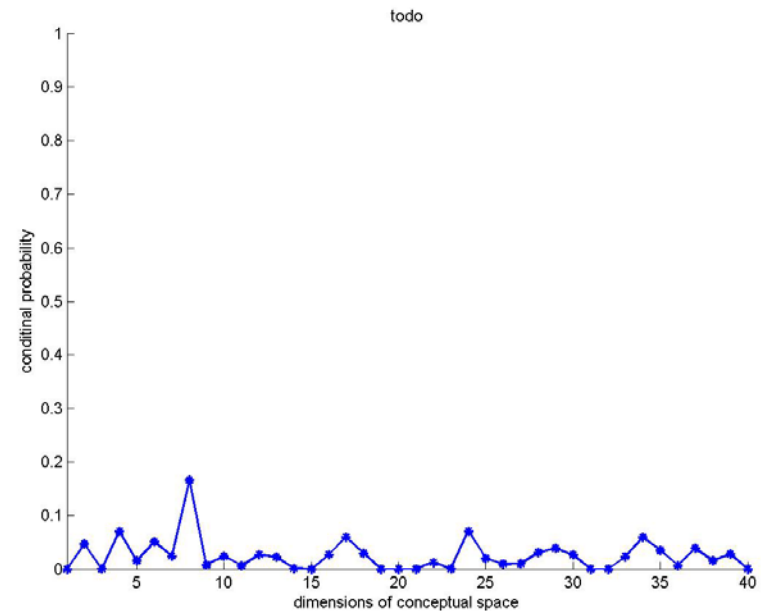
1	java programming Java eclipse software
2	css CSS web design webdesign
3	blog blogs design weblogs weblog
4	music mp3 audio Music copyright
5	search google web Google tools
6	python programming Python web software
7	rss RSS blog syndication blogs
8	games fun flash game Games
9	gtd productivity GTD lifehacks organization
10	programming perl development books Programming

# Fuzziest VS most Definite tag

cooking



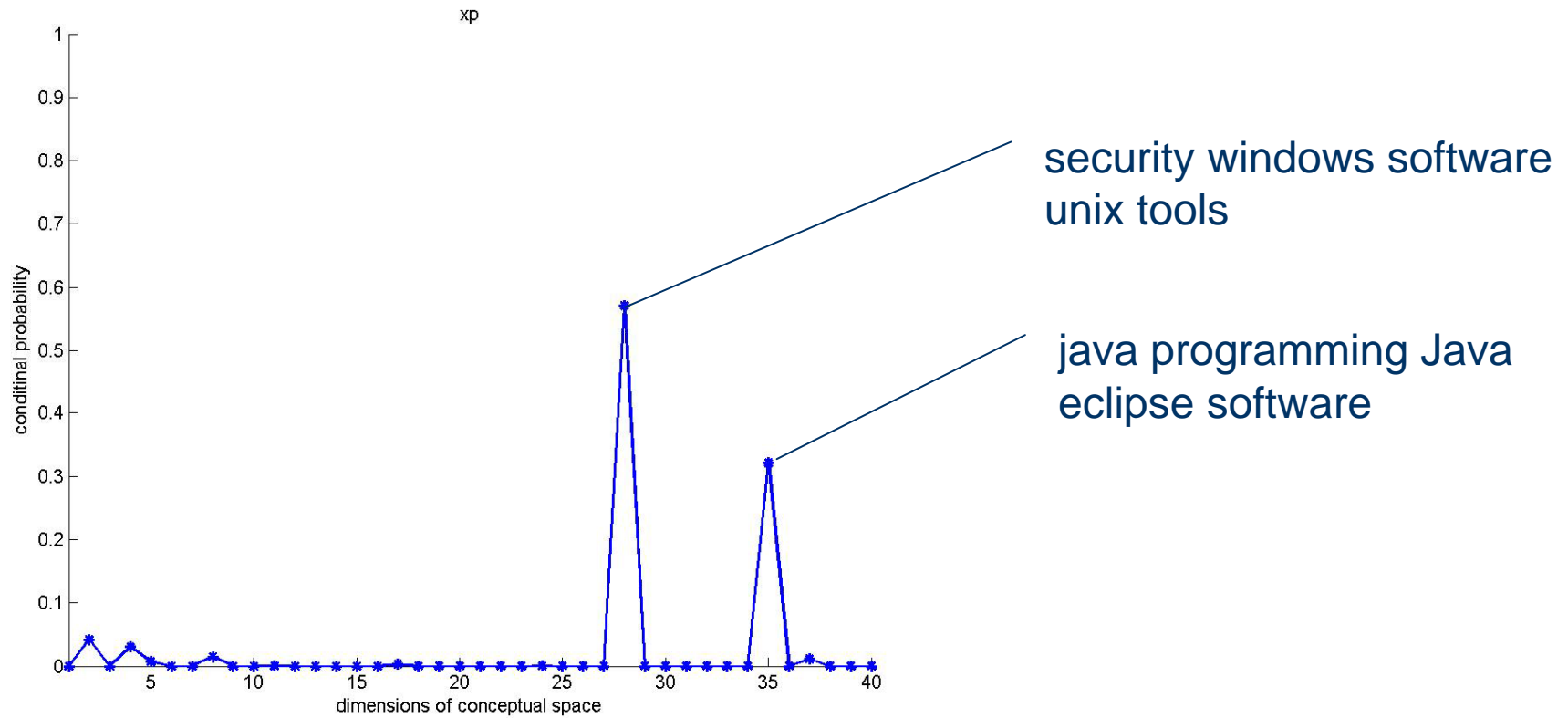
todo



The figures presents the distribution of tags' components over dimensions

By calculating their entropy, we found that cooking is the most definite tag while todo is the fuzziest tag

# Disambiguation of Tag



The vector representation of the tag 'XP' identifies its meaning very clearly through its vector value in the conceptual space.

- ❖ **Background & Motivation**
- ❖ **Exploring Social Annotation**
- ❖ **Semantic Search and Discovery**
- ❖ **Evaluation**

# Basic Search Model

- ❖ The basic model deals with queries that are a single tag and rank semantic related resources without considering personalized information of the user. This problem can be converted to a probability problem.

$$p(r | t) = \sum_{\alpha=1}^D p(r | d_{\alpha}) p(d_{\alpha} | t)$$

# Knowledge Discovery

- ❖ The basic search is thus totally based on the emergent semantics of social annotations without using any keyword matching methods.
- ❖ We can discover related resources by filtering which have been tagged by the query

$$p(r | t) = \begin{cases} \sum_{\alpha=1}^D p(r | d) p(d | t) & : n_{tr} > 0 \\ 0 & : n_{tr} = 0 \end{cases}$$

# Discovery Results for 'delicious'

1	<a href="http://www.betaversion.org/stefano/linotype/news/57">http://www.betaversion.org/stefano/linotype/news/57</a>
2	<a href="http://www.amk.ca/talks/2003-03/">http://www.amk.ca/talks/2003-03/</a>
3	<a href="http://www.ldodds.com/foaf/foaf-a-matic.html">http://www.ldodds.com/foaf/foaf-a-matic.html</a>
4	<a href="http://www.foaf-project.org/">http://www.foaf-project.org/</a>
5	<a href="http://gmpg.org/xfn/">http://gmpg.org/xfn/</a>
6	<a href="http://www.ilrt.bris.ac.uk/discovery/rdf/resources/">http://www.ilrt.bris.ac.uk/discovery/rdf/resources/</a>
7	<a href="http://xml.mfd-consult.dk/foaf/explorer/">http://xml.mfd-consult.dk/foaf/explorer/</a>
8	<a href="http://xmlns.com/foaf/0.1/">http://xmlns.com/foaf/0.1/</a>
9	<a href="http://simile.mit.edu/welkin/">http://simile.mit.edu/welkin/</a>
10	<a href="http://www.xml.com/pub/a/2004/09/01/hack-congress.html">http://www.xml.com/pub/a/2004/09/01/hack-congress.html</a>

# Personalized Search

- ❖ The users interests can be reflects by the websites he tagged and the keyword he used to tag.
- ❖ We can integrate personalized information in the semantic search with the derived emergent semantics.

$$\begin{aligned} p( r | u, t ) &= \sum_{\alpha}^D p( r | d_{\alpha} ) p( d_{\alpha} | u, t ) \\ &= \sum_{\alpha}^D p( r | d_{\alpha} ) \frac{p( u, t | d_{\alpha} ) p( d_{\alpha} )}{p( u, t )} \\ &\propto \sum_{\alpha}^D p( r | d_{\alpha} ) p( u | d_{\alpha} ) p( t | d_{\alpha} ) p( d_{\alpha} ) \end{aligned}$$

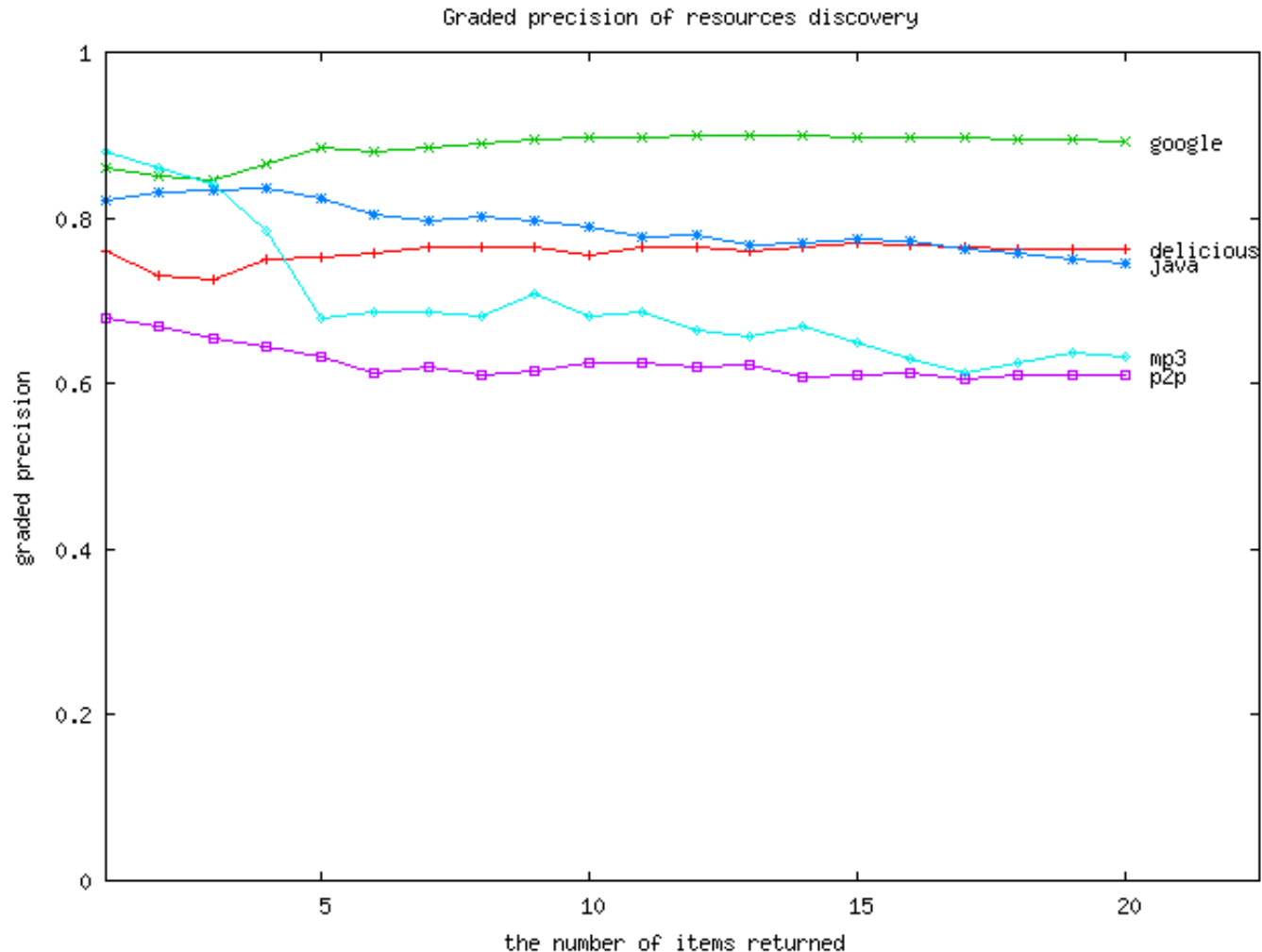
- ❖ **Background & Motivation**
- ❖ **Exploring Social Annotation**
- ❖ **Semantic Search and Discovery**
- ❖ **Evaluation**

# Evaluation Metrics

- ❖ One important difference of our search model is the ability to discover semantically-related web resources from emergent semantics, even if the web resource is not tagged by the query tags.
- ❖ We send the discovery results of 5 widely used tags 'google', 'delicious', 'java', 'p2p' and 'mp3' to people who has computer science background to score.

# Evaluation Results

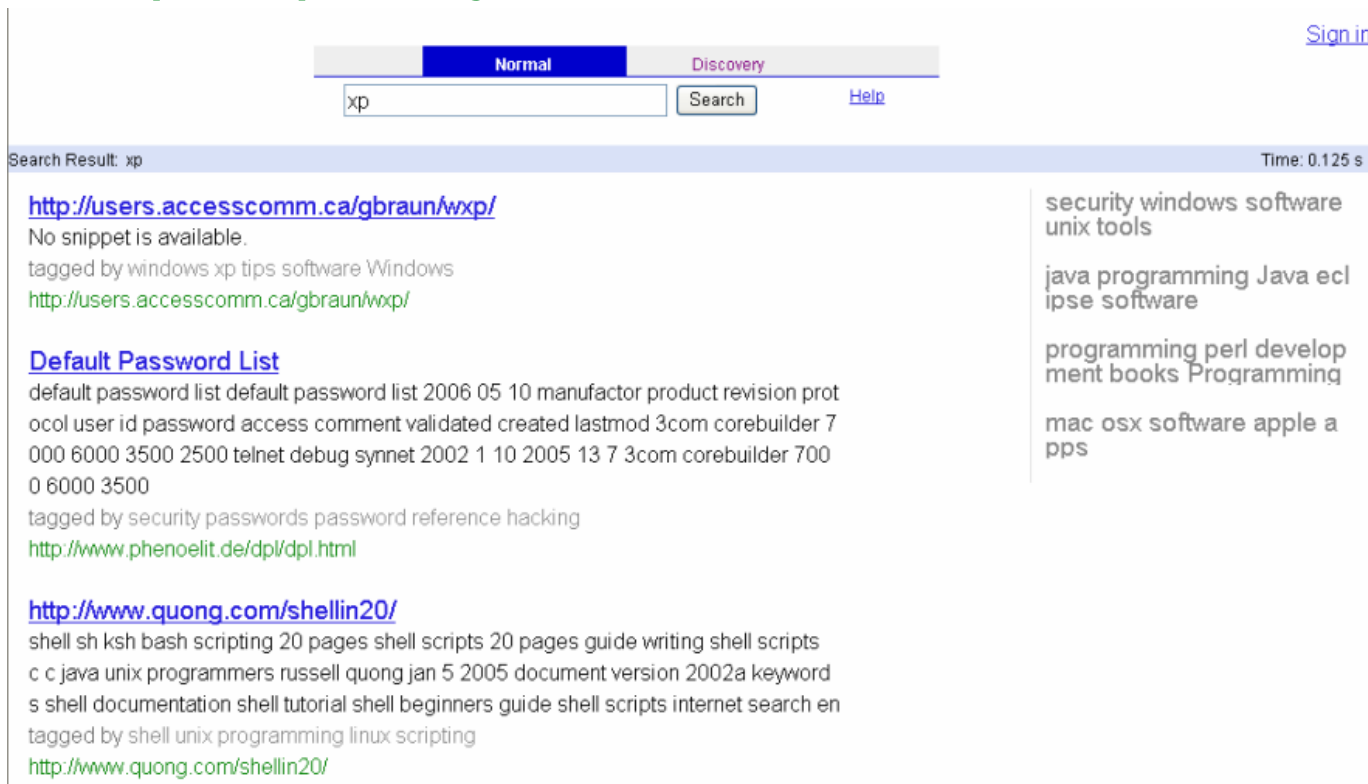
❖ The following figure is the graded precision:



# Demo Application

- ❖ The demo social bookmark search engine can be accessed via:

<http://apex.sjtu.edu.cn:50188>



The screenshot shows a web interface for a social bookmark search engine. At the top right, there is a "Sign in" link. Below it, there are two tabs: "Normal" (selected) and "Discovery". A search input field contains the text "xp" and a "Search" button is next to it. A "Help" link is also visible. Below the search bar, a status bar shows "Search Result: xp" on the left and "Time: 0.125 s" on the right. The main content area displays search results for "xp". The first result is a link to <http://users.accesscomm.ca/gbraun/wxp/> with the text "No snippet is available." and tags "tagged by windows xp tips software Windows". The second result is a link to <http://www.phenoelit.de/dpl/dpl.html> with the title "Default Password List" and a snippet of text about a default password list. The third result is a link to <http://www.quong.com/shellin20/> with a snippet about shell scripts. On the right side of the search results, there is a sidebar with a list of tags: "security windows software unix tools", "java programming Java eclipse software", "programming perl development books Programming", and "mac osx software apple apps".



**Thank You!**