

Abstract

Indexing, both automated and human, has been an important topic for information scientists. Quality indexing is a key to precise and relevant search results. The Internet provides access to massive amounts of data that is cultivating by the second. This phenomenon of information overload provides a strong foundation for granting attention to document indexing.

This study has been inspired by this very topic and is seeking to contribute to our understanding of human indexing on the Internet, especially as we are entering a new era of the Web - one that fosters participation and online collaboration among various communities of interests. Some like to call this new era Web 2.0, where companies are facing the challenge of catering to an environment that is more user-driven and customer-dictated. Enhancing information retrieval has increasingly become an important pursuit to many organizations and soon to be for all.

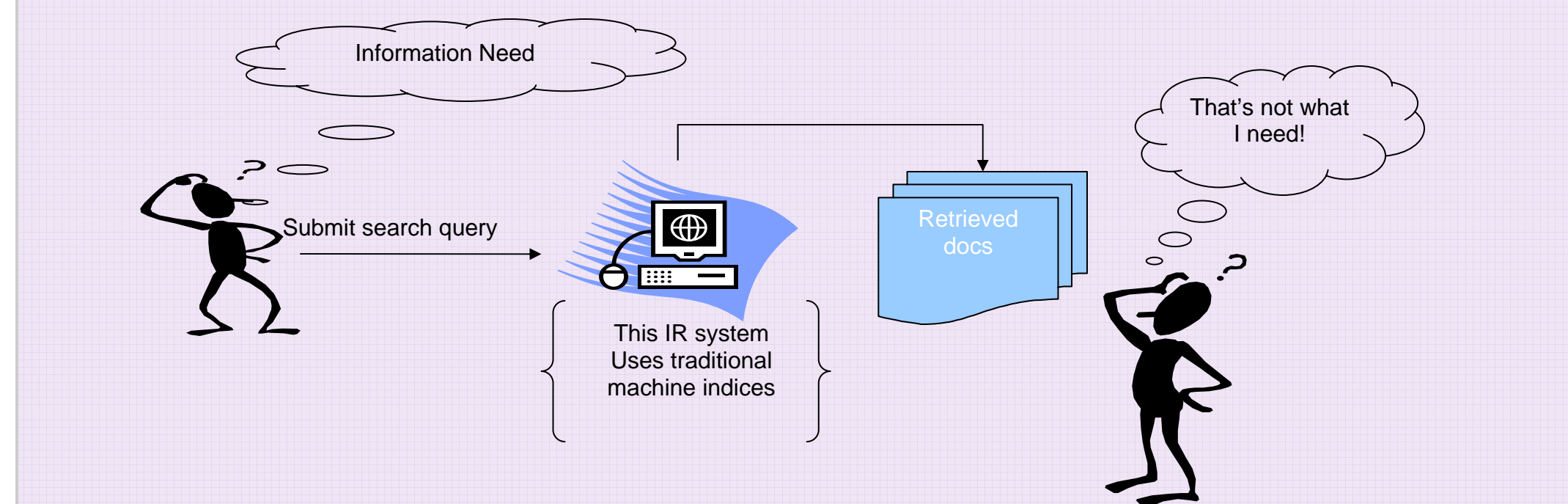
In this project we perform a discovery of human indexing as presented on del.icio.us site. Del.icio.us is a social bookmarking service that is made possible through the use of web-based social software. Del.icio.us employs a user-initiated non-hierarchical keyword categorization model. Registered users tag websites that they find interesting, with as many one-word tags as necessary, for the sake of sharing these websites with others. Del.icio.us was launched in 2003 and is known to be one of the largest social network online.

This study is hoping to provide a human thesaurus that could be used in automated query expansion and translation models to enhance information retrieval on similar collections.

Problem Statement

The inconsistency between document words and search terms submitted by humans makes one of the most fundamental problems in the field of Information Retrieval (IR). This mismatch in indices affect the quality and performance of Information Retrieval Systems (IRS).

This research project is focused on this very problem by attempting to uncover the mismatch between human and machine tags. **Our approach** is different and unique since it is using thousands of real human tags that reflect how humans think and index documents.



Significance of This Work

It is important to understand the nature of human tagging. The more we know about what terms humans are likely to use to search for information, the closer we are to building an IR system that could satisfy the information seeker need efficiently and effectively. Machine indexes have their limitations and may not be able to match the document with the user's search query terms.

This study is attempting to understand human tagging as presented through online social networks where human indices are propagated and made popular when more users decide to use them as tags.

The significance of this study lies primarily in its approach as being one of the early studies that is using real human tags to uncover the semantic hidden relationships between human generated tags and machine tags. Once those relationships are extracted, they could be utilized in query expansion models to enhance IR performance..

Relationship Extracting Method

Input and Output

Input:

- ◆ A collection of documents with paired human and machine tags.
- ◆ The frequency of a machine tag is the occurrence frequency of the tag in a doc.
- ◆ The frequency of a human tag is the number of users indexing a doc with the tag.

Output:

- ◆ Hidden semantic relationships between a human tag and a machine tag

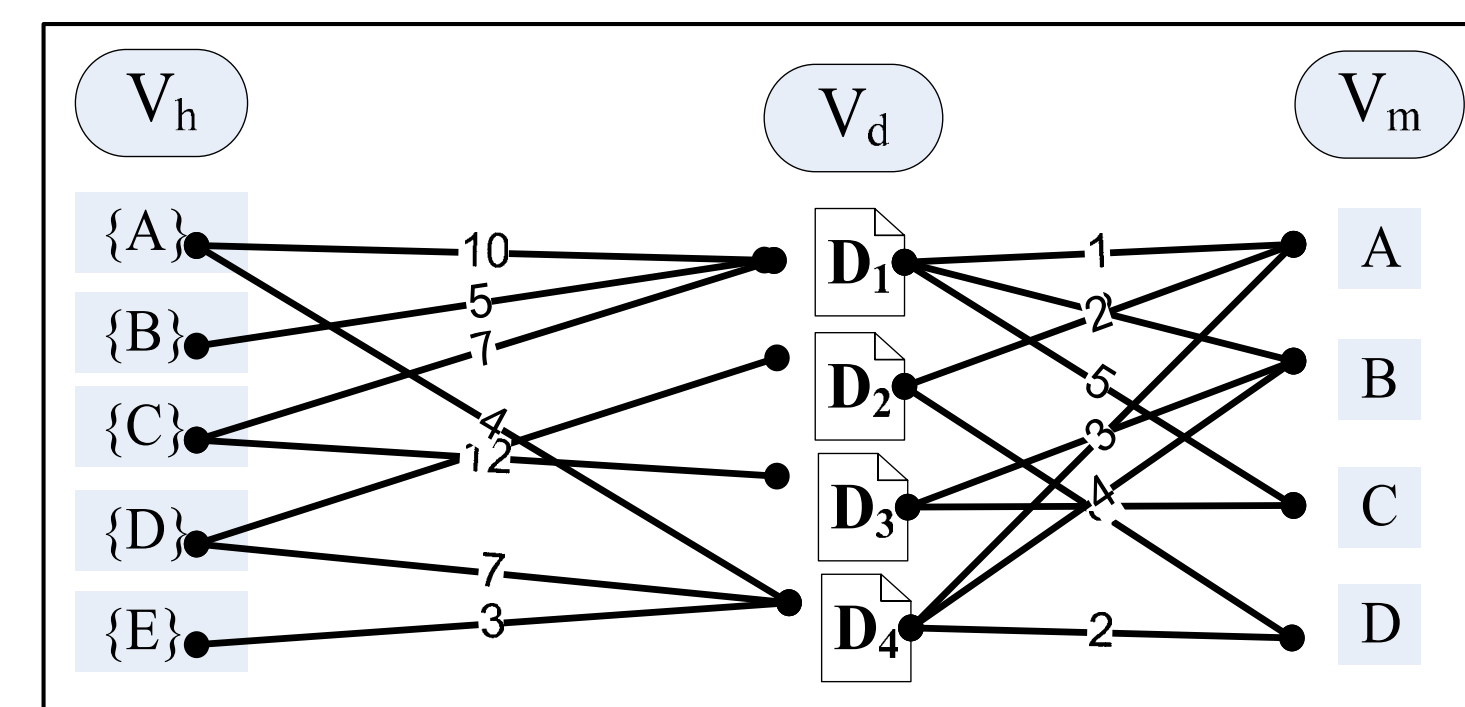


Figure 1. Illustration of document tagging. V_h , V_d , and V_m are human tag set, document set and machine tag set, respectively. The numbers are the frequency of tags in each document.

Human Tagging Assumptions

- ◆ **Correspondence:** Any human tag is based on a related machine tag though they could be different.
- ◆ **Diversity:** For any hidden semantic relationship such as "howto-tutorial", some human indexers may prefer the human tag ("howto"), some may still use the machine tag ("tutorial") to tag the document.

Calculation of Highly Co-occurred Human Tags

Rationale:

- ◆ According to the diversity assumption, if A-B is a hidden semantic relationship, A and B may highly co-occurred in human tagging.
- ◆ Given a human tag, get top N_d co-occurred human tag list, referred to as S_d .

Example: *howto* ($N_d=10$)

{tutorial, reference, tips, software, linux, tutorials, design, photography, tools, windows}

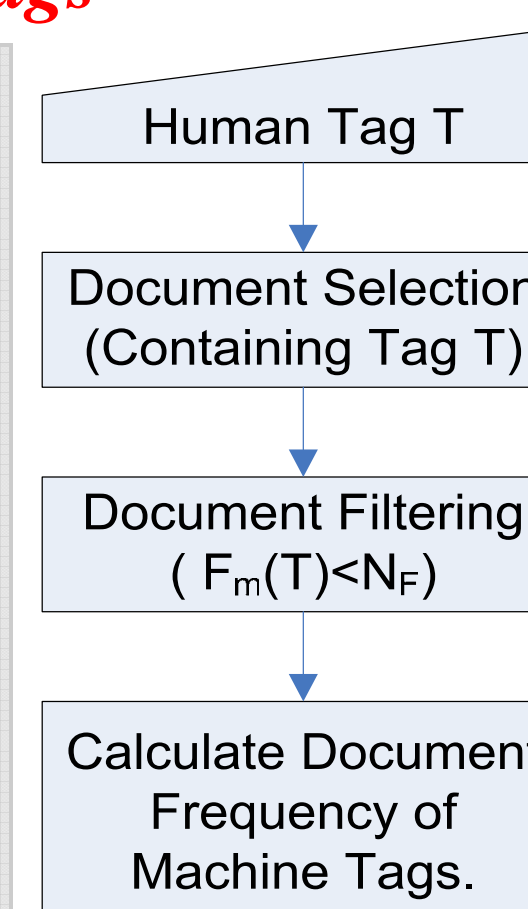
Calculation of Highly Corresponded Machine Tags

Input & Output:

- ◆ Given a human tag T , get top N_c corresponded machine tag list, referred to as S_c .

Method:

- ◆ Only consider document containing T and its corresponding machine tag frequency is less than a small threshold value
- ◆ The machine tags with high document frequency are very likely to be the corresponding hidden machine tag of T .



Final Extraction

Given a human tag, the hidden machine tag which corresponds to the human tag in semantics, is very likely to appear in:

- ◆ S_c according to the correspondence assumption
- ◆ S_d according to the diversity assumption

So, the machine tags appearing in both S_c and S_d are good candidates.

Background

Rowley (1988, p. 43) explained that "the Indexing process creates a description of a document or information, usually in some recognized and accepted style or format" where the term "document" is used to reflect a container of information or knowledge. Therefore, a document could take the form of any combination of form and medium. Document indices could also be viewed as a structured form of content representation of a document. This leads to the notion that indexing is actually the activity of creating surrogates of documents that summarizes their contents (Fidel, 1994).

The process of indexing has two components:

1. Content analysis that results in the selection of the concepts to represent the document; and
2. Translation that is, expressing the concepts selected in the index language used by the information system or database.

Both of these components have long been apart of an active discussion in the IR community on the two fronts of machine indexing and human indexing. While machine indexing has its limitation in capturing the contextual aspects of information seeking request - thus returns a limited set of relevant documents. Human indexing on the other hand, requires rigorous policies and training to provide high indexing quality, which was and still is difficult to achieve. Therefore more attention is being paid to the match between the document index and search terms, which will serve as the foundation of providing relevant set of documents. With the new era of the Internet and the growing popularity of online social software, ordinary people are becoming human indexers with no policies or training. This calls for the need to focus on human-centered indexing and understanding the current search behavior as it stands today.

About del.icio.us

Del.icio.us was developed by Joshua Schachter, who provided a simple interface that allows any Internet users to register and use the service of tagging favorites for free. Del.icio.us became popular among the websphere community as it supported and promoted the use of human tags.

What is a Tag?

A tag is a one-word descriptor used to best describe a favorite. Tags are free of many restrictions, spaces and quotations are prohibited. However, a user could select as many words as desired to tag a favorite. Most people use del.icio.us as a way to organize personal data and share interesting web pages with other users.

Why del.icio.us?

Del.icio.us has an unconventional domain name known as a domain hack. This unconventional name contributed to its popularity, especially since del.icio.us was the second Internet company hosted on a domain hack acquired by Yahoo! after blogs, which was the first.

Social networks and social software is growing in popularity among online communities. Flickr, YouTube and Digg, are three other sites that support social tagging for photos, video and news respectively. Those sites among many more will shape the future of the web.

Interesting Observations:

Del.icio.us tags are suppose to be one-word descriptors where no spaces or quotations are accepted. Thousand of users created and re-used the tag "howto", which is a two-word descriptor if written properly. The pilot study focused on this human created tag, where interesting discoveries were made.

Pilot Study

A pilot study was conducted on the popular human tag "howto" on del.icio.us site. The study revealed that there was a total of 25 popular documents (websites) that people tagged with "howto". None of these documents contained the word "tutorial" in the title and only two documents contained "how to" in the title.

The 25 documents were tagged by a total of 17,960 users who believed that "howto" is a good one-word descriptor of the documents content. The mean for the number of unique users (taggers) per document is 718, while the standard deviation is 1,008.

Table 1. below shows the top 25 documents for tag "howto" and the word frequency of three terms "howto", "how to" and "tutorial". Note that document No. 1, confirms the absence of all three terms in the document, despite the fact that 3,820 unique users believed that the tag "howto" is a good descriptor term. In this case, it is believed that there are some hidden relationships between the human tag "howto" and other document terms that need to be identified.

Table 1. Term presence in top 25 documents with human tag "howto".

Doc No.	No. of Human Taggers	Howto	How To	Tutorial
1	3820	0	0	0
2	306	0	0	0
3	797	0	0	6
4	410	0	0	0
5	236	0	2	0
6	231	2	3	0
7	237	0	0	0
8	916	0	0	0
9	160	0	0	0
10	217	2	0	36
11	3800	0	5	0
12	71	0	0	0
13	195	0	0	0
14	152	0	0	0
15	1658	0	0	0
16	59	0	0	0
17	58	1	3	0
18	283	1	1	4
19	101	0	0	0
20	657	0	2	1
21	405	0	3	0
22	591	0	0	0
23	627	0	0	0
24	1229	0	0	0
25	744	1	1	0

Source: <http://del.icio.us/popular/howto>

Research Questions:

- ◆ How do regular people tag documents?
- ◆ What terms do they select as indices?
- ◆ Do these terms exist in the original document?



Future Work

Future work will include

- ◆ Automating the extraction of the hidden relationships between machine and human tags.
- ◆ Create a measure that represents the confidence of the extracted relationships.
- ◆ Construct a human thesaurus of related machine and human tags to be used for query expansion.
- ◆ Test the effectiveness of information retrieval after using the human thesaurus in query expansion on new collections.

References

- Rowley, J. E. (1988). Abstracting and Indexing (2nd ed.). London: Clive Bingley.
- Fidel, R. (1994). Human-Centered Indexing. Journal of the American Society of Information Science. 45(8): 572-578.
- Anderson, J. D., Perez-Carballo, J. (2001). The Nature of Indexing: How Humans and Machines Analyze Messages and Texts for Retrieval. Part I: Research and the Nature of Human Indexing. Information Processing and Management 37: 231-254.
- Anderson, J. D., Perez-Carballo, J. (2001). The Nature of Indexing: How Humans and Machines Analyze Messages and Texts for Retrieval. Part II: Machine Indexing, and the Allocation of Human Versus Machine Effort. Information Processing and Management 37: 255-277.