A Web-based Tagging Tool for Organizing Personal Documents on PCs

Ji-Lung Hsieh
Department of Computer Science,
National Chiao Tung University
1001 Ta-Hsueh Road, Hsinchu, 300 Taiwan, ROC
gis93813@cis.nctu.edu.tw

Chien-Hsun Chen, I-Wen Lin, Chuen-Tsai Sun
Department of Computer Science,
National Chiao Tung University
1001 Ta-Hsueh Road, Hsinchu, 300 Taiwan, ROC
{gis93582, gis94503, ctsun}@cis.nctu.edu.tw

ABSTRACT
Most desktop computer operating systems provide hierarchical folder mechanisms for managing electronic content, therefore hierarchies still dominate digital information management on PCs. Under certain circumstances, the properties of a folder system make it hard to locate and find specific files stored deep within a hierarchy, therefore researchers are currently studying new mechanisms for file storage, organization, and retrieval. Recently, non-exclusive and flat-network tagging mechanisms have gained popularity for managing online information and files. In this paper we discuss memory theories from cognitive psychology and their potential use in the form of tagging mechanisms for storing and retrieving personal information. Next, we propose a web-based file-tagging tool that allows users to manage files stored on their PCs from remote computers or mobile devices and describe an experiment designed to compare it with a traditional hierarchical folder system.

Author Keywords
Tag, personal information management, document management, human memory, cognitive psychology..

ACM Classification Keywords
H5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

INTRODUCTION
Issues pertaining to the efficient storage, management, and retrieval of files, documents, and other content stored on PCs reflect new demands on human-computer interface design. Using declarative keywords is a common method for organizing content for navigation, filtering, or searches [9]. A desktop search engine like *Google Desktop Search (GDS)* has been recognized as an efficient tool for finding target content in conventional PCs. However, there are at least two conditions under which users cannot rely on search engines to efficiently access files: when users aren't sure they own a copy of a file, and when information is not text-based. Under these circumstances, users can benefit from a systematic information management system for file storage.

Two opposite information management architectures can be utilized by PC operating system. One uses conventional exclusive folders arranged in a hierarchy to categorize files [10], the other describes and filters files using a tagging mechanism. Examples of the hierarchical taxonomy approach include the Linnaean system of classifying living organisms and the Dewey system for library materials [9]. Taxonomy classification is well suited to organizing content along a general-to-specific category continuum, with domain experts establishing classification criteria. Examples of this approach for organizing personal information on PCs include the "My Favorites" feature of *Internet Explorer* and the "Bookmarks" feature of the *Firefox* and *Safari* web browsers for organizing websites of interest to users.

An important characteristic of these tools is that they allow users to create their own classification strategies, although such freedom can be problematic when dealing with files having multiple characteristics that match more than one criterion. For example, an article entitled" Using a social network-based model to simulate flu spreading" can be stored as *c:\paper\social_network\epidemic_simulation* or *c:\paper\epidemic_ simulation\social_network*. If a user places the *epidemic_simulation* and *social_network* folders at the same level but only places the article in one of the two folders, the exclusive property of the other folder may prevent (or at least delay) the user from finding it later [9, 10]. Making two copies and placing one in each folder will eventually cause problems in terms of redundancy and version control when the user wants to update or modify the original. Another problem tied to the general-to-specific approach is that modifications to upper-level general categories influence all lower-level categories.

CHI 2008, April 5–10, 2008, Florence, Italy.

Another disadvantage of organizing personal files within folder hierarchies is that users must remember their classification criteria in the form of keywords. When a file is placed into a sub-folder, a path of multiple keywords is required to properly describe file properties [9]. For example, an article placed in *c:\paper\epidemic_simulation\modeling\social_network\small-world-properties* requires four keywords, a task that becomes more difficult as the number of hierarchy levels increases [9]. The task is made more difficult when users forget their classification strategies.

Tags are non-exclusive and non-hierarchical words that describe files. Tagging mechanisms consist of multiple tags or keywords that describe file content and that can be used for filtering files during searches. The relationship between tags and content can be viewed as a bipartite network, and the relationship among tags can be visualized as a flat complex network with weighted nodes and links. Since multiple files can share identical tags, a content-to-content network can be used to define similarities and relationships [15]. Since a single file can have multiple tags, a tag-to-tag network can be used to describe a personal knowledge ontology (similar to a semantic web ontology [12]) when building a tagging system. The two networks can inform each other and provide different perspectives for users to search, retrieve, and organize their personal knowledge. Since tagging mechanisms are non-exclusive and non-hierarchical, a file does not exclusively belong to one folder but to a group of keywords.

Until 2006, tagging mechanisms were mostly used to organize online documents; popular examples include *Flickr* [7] and *del.icio.us*.[6] Several tagging tools for organizing and searching personal content on PCs have recently been introduced. *Vista*, the latest Windows OS from Microsoft, allows users to tag their files. *VistaGlance* [14] and *Quintura* [11] are third-party products capable of visualizing results from several search engines using keyword networks. *Quintura* is also capable of visualizing search results from *Google Desktop Search* (*GDS*, an application for indexing and searching all files on a PC), but it only allows results to be visualized in terms of keywords and does not allow users to create and organize their own keywords. A tool named *Tagg* [13] gives users the ability to tag personal files and provides network-based visualization. It also allows for the use of tags for file filtering and searching.

The rest of this paper is organized as follows: in the next section we will describe aspects of cognitive psychology as they pertain to tagging mechanisms. In the third section we will introduce a text-based tagging tool named *Personal Knowledge Re-organization by Tagging (PKROT)* for use with PCs, and in the fourth section we will describe an experiment that we designed to test it. In the final section we will discuss our results and offer a conclusion.

COGNITIVE PSYCHOLOGY AND TAGGING

Large amounts of storage, fast processing speeds, convenient graphical user interfaces, and other tools are allowing modern computers to store enormous amounts of information. We believe that information retrievals can be made easier by matching file management architectures with human cognitive memory structures [8]. In the field of cognitive psychology, human memory concepts that are most closely related to this idea are strength and decay, re-processing, and structure. The three-store model of memory offers three hypothetical constructs: sensory, short-term, and long-term storage [2]. Sensory storage is the smallest and quickest to fade, followed by short-term and long-term storage. Memory in terms of computer users is strongly tied to the number of times a specific file is sought, accessed, and updated—that is, the more a user frequents a file, the greater its chances of becoming part of the user's long-term memory. The ease with which files are found depends on a combination of short and long-term memory and file organization.

In contrast to the hierarchy model, Collins and Loftus [4] have proposed a spreading activation model of semantic processing. The non-hierarchical structure of semantic processing is similar to that of flat-structure tag networks. In this model, nodes represent stored concepts but links do not represent subclass relationships as they do in hierarchical models. Nodes can be active or inactive (Fig. 1). Link retrieval (activation) initiates the partial retrieval of related nodes. Stimuli (e.g., from sense organs or adjacent active nodes) trigger relevant inactive concept nodes. Upon activation of a group of nodes, a stimulus spreads throughout inactive nodes until it reaches limitations tied to stimulus energy or active node status.

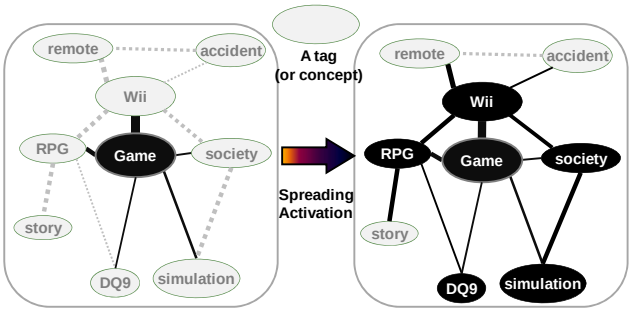


Figure 1. Spreading activation. Node retrieval activates neighboring concepts via connecting links.

Anderson [1] added the network features of declarative knowledge and a spreading activation mechanism when establishing his ACT-R memory architecture. Each node is a cognitive (memory) unit. A spreading activation process determines a level of long-term memory activity that is capable of spreading throughout interconnected cognitive units. Decreased activation is also used to explain the memory priming mechanism—that is, people are more likely to remember ideas presented a short time ago, and the probability of successfully retrieving an idea decays as time

passes. Since most data stored in computers are declarative concepts, user-defined tags can be viewed as cognitive nodes in a concept network and connections between nodes can be viewed as links. We used Anderson's *declarative concept network* as part of our approach to tag-based memory structure and processing.

The levels-of-processing framework proposed by Craik and Lockhart [5, 3] views memory re-processing as the main factor in the enhancement of memory depth and intensity. Memory intensity is heavily dependent on information structure. The more that information is re-processed, the deeper it is stored in human memory and the easier it is to retrieve. These features can be incorporated into the design of a tagging mechanism. By combining the features of levels-of-processing and declarative concept networks, the structure of how users store information influences retrieval probability; by combining the features of the levels-of-processing framework and three-store memory model, information that is most often accessed by users is retrieved more easily. Accordingly, nodes in a tagging network should have different intensity levels depending on re-processing time, and links between pairs of nodes should have different intensity levels according to the number of files they share.

METHOD

We incorporated the three memory frameworks described above in our tagging mechanism as a means of imitating human memory. Our proposed system consists of four parts:

- A method for recording tagged information. For this part we adopted the approach used by online collaborative tagging services so that users can create new tags or use existing tags for each file. Each tag has two scores: association weight and frequency weight. Association weight scores reflect the number of files that share the same tag, and frequency weight scores reflect the number of times a tag has been updated and visited.
- A declarative concept tagging network for storing personal information. The structure or intensity (weight) of nodes and links in this underlying network is modified each time a tag is created, updated, or deleted. We use a list data structure to record association and frequency weights for each tag and frequency weights for each file. Bipartite mapping is used to derive tag-to-tag and content-to-content networks, both implemented using a weight-adjacent matrix data structure.
- A method for updating tagging networks in terms of file reprocessing and intensity decay. Our updating strategy corresponds to a levels-of-processing framework that uses reprocessing as the main factor for increasing memory intensity. User behaviors such as creating, modifying, deleting, reading, and searching for tags that mark a file all serve to increase that file's frequency weight. Frequency weight decay occurs when a file or tag is not utilized for an extended time period.
- A method for retrieving content on demand. The ideal search method for a tag-to-tag network would compare the weighted structure of query keywords to the underlying weighted tag network. However, user interfaces in current use make it difficult to implement this idea. We therefore used the general search engine approach of treating keywords entered by users as filters to retrieve information from tagged networks.

We implemented our system using the high-level *PHP* programming language and a web connection via an *Apache* web server. Users can connect to the web server to upload files, add and organize tags, and perform keyword queries. We constructed a *MySQL* database to record file identities, names, and tags in the form of object modules. The system user interface (created using a template language) can be separated into a query area, file uploading area, results area, tag list, and an area for presenting search results (Fig. 2). Users submit groups of keywords via the query area; multiple keywords represent intersecting relationships between or among keywords. Query results are shown as individual files, each with its own set of tags. The order of presentation of related tags is determined by their association and frequency weights. Tags retrieved by the spreading activation mechanism are also shown for purposes of extending the original query.

Figure 2. Personal Information Re-Organization System overview. Search dialogue is located below page title. Current tags are keywords already queried by the user. Related tags represent those appearing in the results list in order of frequency weight and associated weight. Tag list on the right hand side contains all tags appearing throughout the system.

Our proposed system's primary functions are tagging and searching. The tagging scenario includes file tag addition, deletion, and updating operations. Users can also choose from a list of existing tags. When using remote devices to connect to their own PCs, users can upload files and add related tags via the Web. In the search scenario, multiple keywords will initially trigger results that represent keyword intersections. Clicking on other file tags in the results frame activates additional conjunctive queries consisting of new and previously used tags. This mechanism is implemented to imitate the spreading

activation model (Fig. 1). Users can also remove previously queried keywords to obtain more general results. As with existing systems such as *Quintura* and *Tagg*, we make it possible to "do and undo" searches by clicking on queried tags.

EXPERIMENT AND ANALYSIS

We designed an experiment to test our proposed tagging mechanism. Participants were 64 third-year students recruited from a senior high school in northern Taiwan. The experimental (tagging mechanism) group consisted of 37 students in a single class and the control group (conventional folder system) consisted of 27 students in another class. Participants were asked to use their respective systems to classify 60 illustrations. Their efforts were saved for one week, after which participants were asked to find four specific pictures. Computer screen recorders were used to monitor student classification schemes and file tags and to observe search strategies. We counted the number of steps and measured how much time each participant spent finding each of the four pictures. We calculated the average number of tags used by experimental group students to determine how well they understood the tagging mechanism. As shown in Figure 3, they used an average of 3.1 tags to describe each file, indicating a thorough understanding of the tagging concept. Finally, data on computer/Internet usage and frequency were collected by means of a survey administered at the end of the experiment.

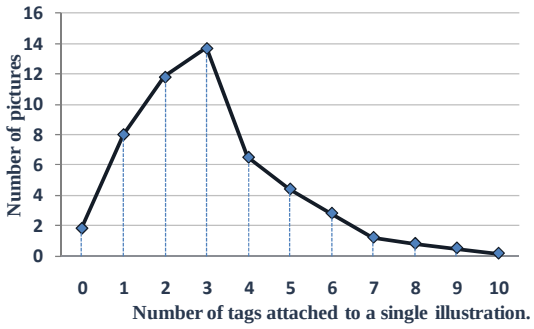


Figure 3. Average number of tags attached to individual illustrations. Over four-fifths (83%) had at least two tags.

To prevent students from favouring certain illustrations, our descriptions of and instructions for finding the four pictures were purposefully written in an ambiguous manner. The basic instruction was simply to find the four files (each containing a picture expressing multiple concepts) as quickly as possible. For example, we asked students to find a picture of an "unreal cat"; one of the 60 pictures was of the famous cartoon character named Garfield. We observed that some students in the control group placed their Garfield pictures in a folder labelled "cartoon," while several students in the experimental group used the tags "cartoon" and "cat." Under such circumstances, we predicted that participants who used the tagging system, especially those who used multiple tags, would be faster in completing the assignment. Another example is our inclusion of Monet's "Impression Sunrise," whose composition contains a small ship. Participants were asked to locate an image of "a little ship." To maintain consistency in the folder and tagging classification environments, we created a system that allowed students in both groups to preview thumbnails of the illustrations.

As shown in Figure 4, control group students needed more time than experimental group students to find the four pictures, regardless of how they organized their folder hierarchies. However, the data on the number of "clicks" required to locate the illustrations indicate that the experimental group students consistently took more steps to complete the search assignment (Fig. 5). There are at least two possible reasons for this unexpected result: a) the tagging students were more accustomed to using a folder system to organize files, and b) the folder system has an exclusive property [10] that may help users remember their folder organization hierarchies to reduce the number of search steps. Again, despite the higher number of clicks-per-search, the time required to complete searches was shorter for users of the tagging system.

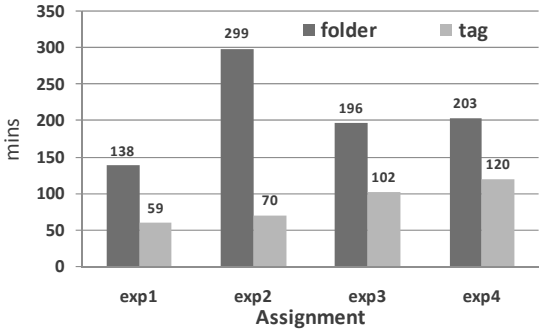


Figure 4. Time spent searching for four illustrations by folder system (control group) and tagging system (experimental group) participants.

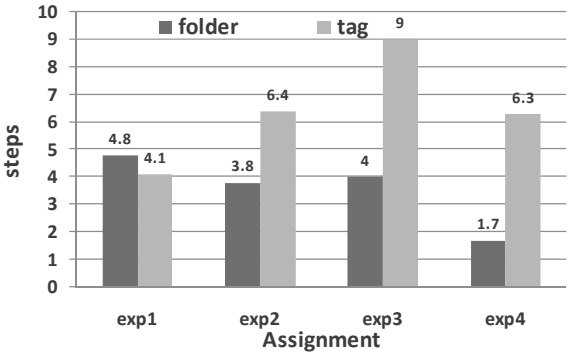
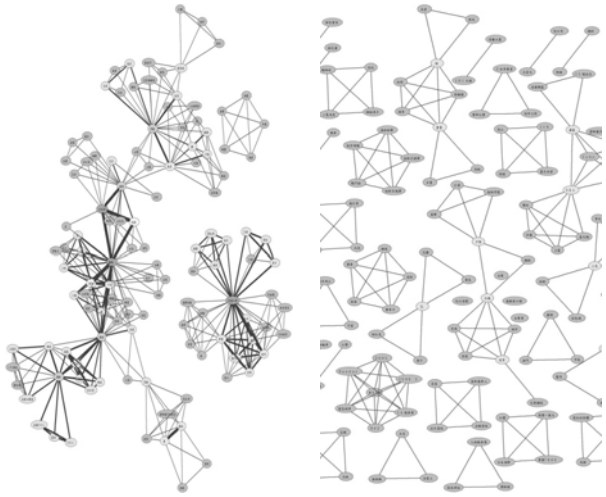


Figure 5. Number of steps taken by folder system (control group) and tagging system (experimental group) participants to complete the illustration search assignment.

We created visual representations of the tagging networks built by the experimental group participants. The structures are similar to those of the declarative concept network described in the second section of this paper, leading us to suggest that tagging networks are structurally similar to human memory processes. We also observed two memory styles. The structure of tagging networks for participants who tended to use existing tags and common concepts to describe picture content resembled a connected plane graph (Fig. 6a). The structure for participants who tended to use specific concepts and new tags resembled a disconnected plane graph (Fig. 6b). In terms of the above-described spreading activation mechanism and declarative concept network structure, a connected tagging network apparently triggers spreading activation more easily, while a disconnected tagging network obstructs spreading activation.



6a 6b

Figure 6. Two tagging network approaches used by the participants. Each node represents a tag. Link thickness reflects the number of files using both tags simultaneously. Figure 5a represents a network with a large number of strong connections.

CONCLUSION

Since our tagging mechanism was informed by some current theories of human memory, we suggest that it is suitable for at least four situations: a) non-text media (e.g., figure files or portable-document format files that are difficult to search for using keywords); b) articles with multiple concepts (especially academic articles or pictures); c) content requiring frequent searches and visits (e.g., academic papers); and d) when dealing with large numbers of files lacking suitable file names. These four conditions are frequently marked by memory loss, ambiguity, and redundancy.

However, the folder hierarchy approach still outperform tagging mechanism in the following classification scenarios: a) professional knowledge such as patents or library books; b) media with fixed properties (e.g., music, which can be categorized by year, composer, and performer); and c) exclusive properties (e.g., jazz singers who seldom, if ever, sing songs from other musical categories).

A final task for designers of a useful tagging mechanism is creating a system that allows users to visualize how they organize or record information. Graphical user interfaces that help users perform searches in tagging networks have been developed by *Quintura* and *Tagg*. These systems often consist of a plug-in for visualizing the tagging network and a main window to show search results. However, since the amount of data that can be stored online or in PCs is now extremely large, coming up with a viable method to visualize large, complex networks remains a challenge for developers and designers. A complex network consisting of thousands of nodes (e.g., a network consisting of over 100,000 patents published in a single year) can easily overwhelm a user. Currently, most graphical user interfaces specifically designed for complex networks are limited to providing directed access. A new approach is required to logically present large amounts of information.

REFERENCES

1. Anderson, J.R. A spreading activation theory of memory. *Journal of Verbal Learning and Verbal Behavior*, 22(1983), 261-295.
2. Atkinson, R.C. and Shiffrin, R.M. *Human memory: A proposed system and its control processes (Vol. 2)*. Academic Press, New York, 1968.
3. Brown, S.C. and Craik, F.I.M. *Encoding and retrieval of information*. The Oxford handbook of memory, New York: Oxford University Press, 2000, 93–108.
4. Collins, A.M. and Loftus, E.F. A spreading activation theory of semantic processing. *Psychological Review*, 82(1975), 407-428.
5. Craik, F.I.M. and Lockhart, R.S. Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior*, 11(1972), 671-684.
6. *del.icio.us*. http://del.icio.us/.
7. *flickr*, http://www.flickr.com/.
8. Furnas, G.W., Fake, C., von Ahn, L., Schachter, J., Golder, S., Fox, K., Davis, M., Marlow, C. and Naaman, M. Why do tagging systems work?. *Ext. Abstracts CHI 2006*, ACM Press, (2006), 36-39.
9. Golder, S. and Huberman, B.A. The Structure of Collaborative Tagging Systems. *Technical report. Information Dynamics Lab, HP Labs*, (2005).
10. Jones, W., Phuwanartnurak, A.J., Gill, R. and Bruce, H. Don't take my folders away! Organizing personal information to get things done. In *Proc. CHI 2005*, (2005), 1505-1508
11. *QuinTura*, http://www.quintura.com/

12. Shirky, C. Ontology is overrated: Categories, Links and Tags.http://www.shirky.com/writing/ontology_overrated .html, (2005).
13. *Tagg*. http://www.taggtool.com/index.php.
14. *VistaGlance*. http://www.vistaglance.com/.
15. Watts, D.J. *Six degrees: the science of a connected age*. W.W. Norton & company, New York, 2003