You Are What You Tag

Yi-Ching Huang and **Chia-Chuan Hung** and **Jane Yung-jen Hsu**
Department of Computer Science and Information Engineering
Graduate Institute of Networking and Multimedia
National Taiwan University
{r95045, r95944001, yjhsu}@csie.ntu.edu.tw

Abstract

People establish *personal profiles* to obtain online services. Profiles consisting of simple factual data provide an inadequate description of the individual, as they are often *incomplete*, mostly *subjective* and cannot reflect dynamic changes. This research explores the idea of "you are what you tag", namely, an individual can be effectively profiled by the tags associated with his/her social media. In particular, this paper presents the *personal*, *social*, and *global* views of a person's profile based on the tags and content of social bookmarking. To facilitate the alternative views, profiles are visualized as tag clouds based on color harmonic combinations. Commonsense semantic analysis and co-occurrence measurement are defined to calculate tag similarity. Therefore, the proposed approach supports an intuitive, natural and novel interface for people to browse/search through a social web site.

Introduction

People establish *personal profiles* at many online communities in order to obtain specific services. For example, on a job search website, job seekers maintain their basic personal data, resumes, skills and interets. Recruiters browse through these profiles to identify qualified candidates for the jobs. A typical personal profile consisting of simple factual data, such as the name, affiliation, or interests, provides an inadequate description of the individual. First of all, due to privacy concerns, most users are reluctant to provide more information than what is required by the service. Secondly, user-specified profiles are mostly *subjective*. In the job search scenario, it is risky to judge whether a person is a good candidate for a job from his/her resume alone. Opinions from a candiate's friends or former employers can be valuable in forming a full picture of the candidate. Comments on strong job performance and great personality from the references can be the key to a positive decision. Lastly, such simple user-specified profiles do not reflect dynamic changes, even though skills and interests of a person do evolve over time.

This paper presents an novel idea of "you are what you tag" for user profiling. Namely, an individual can be effectively profiled by the tags associated with his/her social media. Our basic assumption is that the rich online media produced/consumed by an individual can reveal important features about the person. Many online services today provide a platform for users to publish digital contents, which can be tagged. For example, the photo collections on Flickr show the people, places, and activities engaged by the user; the bookmarks on del.icio.us represent the topics of interest to the user; the blog posts on Blogger reflect the events, social interactions, or feelings experienced in the author's life. The user-specified tags associated with these personal collections of digital contents along with their comments provide meaningful descriptions of a person.

Our research explores tag-based user profiling on several levels. Without loss of generality, this paper presents the *personal*, *social*, and *global* views of an individual's profile based on the tags and content of his/her social bookmarks. In particular, the bookmark data are acquired from del.icio.us, a social bookmarking website.

Instead of attempting to collect the opinions of a group of people directly, we examine the tags assigned by others to the contents in one's collection. When many people tag a person's content collection with a specific keyword, it is natural to assume that the content owner shares the same idea. For example, suppose that everyone tags a specific collection with "movie", it is reasonable to conclude that the owner of the collection is fond of movies. In other words, the collective wisdom shared by most people defines the global view about the person.

To facilitate the alternative views, profiles are visualized as tag clouds based on color harmonic combinations. Commonsense semantic analysis and co-occurrence measurement are defined to calculate *tag similarity*. Therefore, the proposed approach supports an intuitive, natural and novel interface for people to browse/search through a social web site. For a collection of one's bookmarked contents, we switch to different viewpoints by examining other people's tags on this collection. Three types of view are defined: *personal*, *social*, and *global views*. To illustrate the differences between these viewpoints, we design a way of visualization, that is, by grouping tags with similar concepts and displaying them with similar colors, viewers can be more attentive to focus on the information he/she needs.

In the remainder of this paper, we will start by briefly re-

viewing some related work. The proposed data model will be defined next, followed by the methods for analysis from three different views. Visualization of the personal, social and global profiles are then presented before a short discussion and the conclusion.

Related Work

Research on InterestMap (Liu & Maes 2005) harvests profiles from social networking websites, such as Friendster[1], MySpace[2], and Orkut[3], to construct the InterestMap, a network-style to illustrate the relationship between interests and identities. Rather than traditional recommender systems, they recommend by considering the interest of people instead of the historical behavior in a particular application. They try to model people's preferences and interests in real life. This work not only upgrade the accuracy of recommendation but also provide a visual way to explore one's interests on InterestMap.

There are different kinds of social networking websites including flickr[4], del.icio.us[5], last.fm[6] and others. On these social media websites, people are free to tag the multimedia content and share their contents with their friends or the general public. Tagging is a social indexing process and contents can be categorized by any number of tags. As the number of tags increases, it becomes useful to view these tags visually. The *tag cloud* is a visual interface to help people retrieve important information quickly. In (Hasan-Montero & Herrero-Solana 2006), they reduce the semantic density of a tag set and improve the visual consistency of the tag cloud layout. An approach to tag selection was proposed and a clustering algorithm is used to produce visual layout. In (Kaser & Lemire 2007), some models and algorithms to improve the display of tag cloud in HTML were presented.

Different from InterestMap, we utilize the tags in these social media content which people collected. We apply statistical and commonsense reasoning to establish semantic connections among these tags. All tags are grouped by their concepts for visual layout rather than the alphabetical order in a traditional tag cloud. Futhurmore, we measure three different viewpoints to provide a comprehensive representation of a person. People can then compare these three profile models using our visualization function to learn more information about a person.

Tags and Social Network

At a growing number of social media websites, tagging plays an important role of helping user manage their documents. Users are encouraged to add tags to describe a document and to share these tags with other people. These tags indirectly reflect a user's interests, concerned topics, and activities in daily life, etc., thus can serve as the building blocks of a user's profile.

[1]http://www.friendster.com
[2]http://www.myspace.com
[3]http://www.orkut.com
[4]http://www.flickr.com/
[5]http://del.icio.us
[6]http://www.last.fm/

In this paper, we propose using tags to profile a person instead of a typical profile list. The advantages of using tags include:

- the flexiblity in describing a person using any term,
- the precision in presenting a person's preferences, and
- the ease in showing the different views of a person.

The proposed idea can be applied to any type of social media with tags. In this paper, for simplicity, social bookmarks are used as the data source, and each bookmarked document is assumed to have multiple tags.

Automatic construction of social networks from various social media data sources is an important research topic in information mining. There have been several approaches to constructing social networks. For example, social networks can be extracted from blog posts and comments (Furukawa *et al.* 2007). Social networks can also be mined from photo collections provided that the photos are annotated with rich metadata of the people in the photo (Huang & Hsu 2006). In what follows, let's assume that the social network is given.

Data Modeling

We propose to add social media as a new type of nodes into social network. We call this network a *Social Media Network*. For example, Figure 1 represents the common social media between a person and his/her friends. The edges between round-shaped nodes represent the traditional social network, while an oval-shaped node represents a bookmark URL, and a dotted directed arrow means someone has this bookmark. Note that it also shows someone, denoted as P_x, P_y etc, who does not connect with others (meaning P_x, P_y do not know them) but also has this bookmark URL.

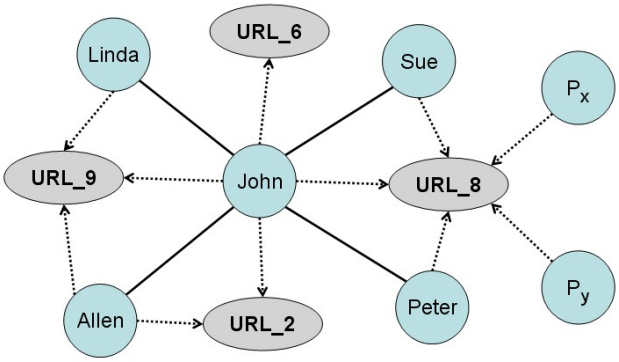

Figure 1: A sample social media network.

Figure 2 illustrates the tags given by a user to his/her bookmark URLs. A document is tagged with multiple tags, which help to describe the document in different *capacities*. For example, the tag "map" has a capacity calculated as 0.9 to describe URL_7. Furthermore, a user may tag different documents with the same tag. Thus for each tag, we give the attribute *strength* to symbolize the importance of the tag for its owner. In the next section, we describe how these tag weights are determined.

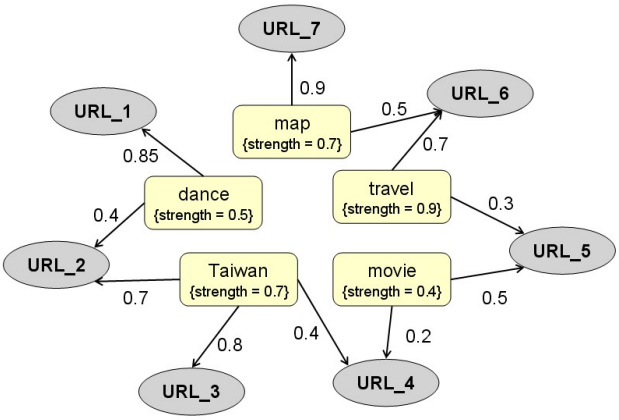

Figure 2: An example of a person's tags and his/her own bookmarks. For each arrow, the floating number denotes the capacity of this tag to describe this document. Each tag has the strength attribute to represent the importance of the tag for its owner.

Tag Analysis

Each bookmarked URL is given one or more tags to describe the content of the webpage. We define a value, *capacity*, to represent how much a tag can describe the content of a document. By analyzing the nature and idea of a tag, the frequency of assigning an identical tag to the same content, and the tagging order, we can determine the capacity of a tag. The importance of a tag to an owner, denoted as tag *strength*, is determined by analyzing its capacity, the volume of covered bookmarks, and the quality of its bookmarked content.

Tag Capacity

Naturally, a tag used by more people to describe an identical bookmark link has a higher importance to this bookmarked document. The first tag may be more relevant than the second tag when a user tags this document. A document with diverse semantic tags may indirectly show that each tag only represents a part of the document. Thus we consider tagging frequency, tagging order, and tags diversity when trying to determine the capacity of a tag to describe a bookmarked document.

We denote a collection of bookmarked documents as $\mathbf{D}$, and a set of tags $\mathbf{T}$ which are tagged on $\mathbf{D}$. A tuple of bookmarking data is denoted as $b = (p, bT, d)$ which means a person p who tagged document $d \in \mathbf{D}$ with a sequence of tags $bT = \{bt_1, bt_2, \cdots, bt_n\}$. We first define the order weight of tag $bt_i \in bT$ as

$$w_{order}(bt_i) = \begin{cases} \exp^{-i/10} & \text{if } i \leq 10 \\ \exp^{-1} & \text{if } i > 10 \end{cases} \quad (1)$$

where $bt_i \in bT$ and i is the index of bt_i in this ordered tagging sequence. Here we let tags after the 10^{th} tag have equal order weight.

Second, we group tags in bT with their concept. Each cluster represents a concept in this document. We assume that more concepts on a document will reduce the weight of each concept, and a larger cluster is more important for this document. We use ConceptNet and co-occurrence measurements to group these tags. More details are described in next section. Each tag belongs to only one cluster. Let $bT_c = \{bt_1, bt_2, \cdots, bt_k\}$ be a cluster of bT with k tags, then we define the concept weight of $bt_i \in bT_c$ as

$$w_{concept}(bt_i) = \frac{1}{|bT|} * |bT_c| \quad (2)$$

Note that $w_{concept}(bt_i)$ is higher when bt_i belongs to a larger cluster, and it is lower when there are many clusters. Each tag belonging to a cluster has the same concept weight. The summation of tag number of each cluster equals to the number of tags in bT

Combining both weight measurements, for each $bt_i \in bT$, we can define the weight of bt_i for a document d given by a person p as

$$w_p(bt_i) = \delta * w_{order}(bt_i) + (1 - \delta) * w_{concept}(bt_i) \quad (3)$$

where δ is an argument that can be adjusted.

Finally, we take all the people who tag the same document with the same tag into consideration and sum up the weights provided by everyone, and we can determine the importance of a tag t for a document d. We define $\mathbf{M}$ as a set of people who ever tagged document d with tag t, thus the the capability of t on d is

$$capacity(t, d) = \sum_{p \in \mathbf{M}} w_p(t) \quad (4)$$

Personal Profile

To use tags to present a person, we first need to determine the weight of each tag for this person according to the document which this person has bookmarked. Each tag has different importance in representing a person. For example, if one has more bookmarks on drama than on classical music bookmarks, , it may be concluded that he/she likes drama more than classical music. Thus the tag "drama" should have higher weight value than the tag "classical music". As mentioned previously, each tag has a different capacity in describing a document. We can summarize the weights of an identical tag on one's own documents, and obtain a total weight of a tag for representing this person.

However, not all documents have a high quality. Here we assume that the document bookmarked by many people has higher quality. Given a collection of documents $\mathbf{D}$, N_P is the total number of people who ever tagged any document in $\mathbf{D}$, and $N_{p \to d}$ is the number of people who ever tagged d, then the quality of d is $Q(d) = N_{p \to d}/N_P$.

To examine one's bookmarks, we can do so from the view of this person, thus we can obtain a set of tags that can subjectively present he/she. However, sometimes we are interested in the views of other people as they provide different viewpoints and sometimes more objective facts. In this paper, we define three types of viewpoints to consider one's bookmarks: *personal*, *social*, and *global views*.

Suppose p is the person we wish to view, and $\mathbf{D}_p$ denotes the total documents that p has bookmarked. Our goal is to determine a sets of tags $\mathbf{T}_p$ that can represent p's characteristics. The result $\mathbf{T}_p$ may be different from each viewpoint although we are describing the same person.

Personal viewpoint From the personal viewpoint, we only consider the tags assigned by p, and denote these candidates as a set $\mathbf{CT}_p$. For each candidate tag $t \in \mathbf{CT}_p$, we define the *strength* of t from p's view as

$$strength_p(t) = \sum_{\forall d \in \mathbf{D}_p} capacity(t, d) \quad (5)$$

where subindex p means from p's view. Thus

$$\mathbf{T}_p = \{t | t \in \mathbf{CT}_p \text{ and } strength_p(t) \leq \alpha\} \quad (6)$$

where we set a threshold α to filter those tags with strength lower than threshold. (α is a configuration that can be adjusted by the viewer.)

Social viewpoint From the social viewpoint, we consider the tags assigned by the actors on p's personal social network. Note that we still focus on documents in $\mathbf{D}_p$, and our candidate tags are those tags which were assigned by p's acquaintances on these documents. Let $\mathbf{CT}_s$ denote the set of candidate tags, $\mathbf{A}$ is p's acquaintances, and $\mathbf{A}_t$ denotes a set of acquaintances who have tag t. Thus for each $t \in \mathbf{CT}_s$,

$$strength_{social}(t) = \frac{1}{|\mathbf{A}_t|} * \sum_{a \in \mathbf{A}_t} strength_a(t) \quad (7)$$

And

$$\mathbf{T}_p = \{t | t \in \mathbf{CT}_s \text{ and } strength_{social}(t) \leq \alpha\} \quad (8)$$

Note that we can generally define $\mathbf{A}$ as any group of people, and $\mathbf{CT}_s$ as the tags assigned by people in this group. Suppose we define $\mathbf{A}$ as a set of experts, we can see the opinions from experts for these documents. It can be used to determine how much domain knowledge p has.

Global viewpoint From the global viewpoint, we consider the tags assigned by everyone on $\mathbf{D}_p$. This viewpoint can be thought of as a generalization of the social viewpoint. $\mathbf{A}$ represents the set of people who ever tagged any document in $\mathbf{D}_p$, and all of their tags are elements in candidate tag set $\mathbf{CT}_u$. The results of this viewpoint can reflect the common opinions from public, and probably is most objective.

Visualization

As it is time-consuming for people to understand the meaning from large amounts of data, we design a visual way to show a person's profile. Viewers can compare the difference between personal characteristics of distinct individuals.

To facilitate the alternative views, profiles are visualized as tag clouds based on color harmonic combinations. Commonsense semantic analysis and co-occurrence measurement are defined to calculate tag similarity. Therefore, the proposed approach supports an intuitive, natural and novel interface for people to browse/search through a social web site. For a collection of one's bookmarked contents, we switch to different viewpoints by examining other people's tags on this collection. Three types of view are defined: *personal*, *social*, and *global views*. To illustrate the differences between these viewpoints, we design a way of visualization, that is, by grouping tags with similar concepts and displaying them with similar colors, viewers can be more attentive to focus on the information he/she needs.

A personal profile is composed by tags, therefore we plan to use tag clouds to present the profile. General tag cloud use font size and color to emphasize the frequency of tag usage. Users can interact with tag clouds and browse the detail of tag-described resources when he clicks on a tag. With growing number of tags, alphabetical arrangements of displayed tags are insufficient. When people want to search for a certain type of information, they have to scan all the tags and infer the semantic relationship between the tags. Such visual layout fails to present the complete meaning of the tags and can be extremely time-consuming. We define that similarity-based layout to improve tag clouds. We use common sense to cluster the similar tags and provide two-dimension tag placement to let the user easily know the relationship among tags. People can understand the semantic meaning of the tags and obtain overall characteristics of a person. To improve the quality of visualization, we utilize color harmonic combinations to map out different clusters of tags. Tags in the same cluster use the same hue and different value in color theory; similar clusters use the same color tone to emphasize the relationship among them. The visualization is like Figure 3.

Figure 3: Tag visualization

Semantic Similarity and Co-occurrence

To cluster the tags automatically, we propose two methods for calculating tag similarity and cluster the ones with high similarity. One adopts commonsense reasoning similarity and the other uses relative co-occurrence statistics.

The ConceptNet-based semantic similarity Tag is in fact text and contains semantic meaning. Some tags have similar concept and some have different ones. We plan to use

commonsense reasoning to obtain related tags with a similar concept. In this paper, we utilize ConceptNet to support the content analysis. ConceptNet is a freely available commonsense knowledgebase and it provides a natural-language-processing toolkit for reasoning tasks including "topic-jisting", "analogy-making", and "text summarization".

ConceptNet is a semantic network created by Hugo Liu and Push Singh(Liu & Singh 2004). It collects commonsense knowledge from the Open Mind Common Sense corpus and contains 300,000 nodes and 1.6 millions links, such as (IsA 'apple' 'red fruit') or (PropertyOf 'game' 'fun'). The ConceptNet toolkit provides node-level and document-level reasoning operations. Two functions on textual analysis(Liu & Singh 2004) are introduced:

- GetContext(): It accepts the input of a textual document which is then translated into a ConceptNet-compatible format. It finds the neighboring relevant concepts using spreading activation around this concept of document. For example: the neighborhood of the concept "music" includes "play violin", "play piano", "band", etc.
- GuessConcept(): It takes input as a document and a novel concept in that document, and it outputs a list of potential items which are analogous to the input concept. In other words, it can obtain analogous concepts from the concept of input document. For example: the concept of "do exercise" is analogous to "ride bicycle", "play football", etc.

In this paper, we utilize the two functions mentioned above to measure the semantic distance between two textual words, which we define as "tags". We use a spreading activation algorithm(Collins & Loftus 1975) to conduct its inferences and compute the similarity among tags. First, we use GuessConcept() to acquire a list of analogous tags given by a tag. Second, we use GetContext() to acquire all neighboring relevant tags given by analogous tags using spreading activation. The input tag as a first node has highest level of energy and spreads a fraction of its energy to relevant tags. The value of spreading energy is directly proportional to the weight between tags. The energy of any tag after a spreading step is calculated by Equation (9):

$$t_i = \sum_{j=link_j} energy(t_j) * weight(t_i, t_j) \qquad (9)$$

where t_i is original tag and t_j is a tag activated by t_i. $energy(t_j)$ is energy of t_j acquired from t_i and $weight(t_i, t_j)$ is a link weight between t_i and t_j.The energy of the tag would decrease at a ratio α step by step, and stop until no new tags are activated. Finally, we collect the activated tags which are the relevant tags.

We utilize our visualization approach to present three viewpoints of a person's profile. For instance, the manager of a travel agency in search of a tour guide to fill a vacancy sees a person's profile from three viewpoints as Figure 4, Figure 5, and Figure 6. The manager could obtain more information such as "map"or "trip" that they otherwise infer simply from personal viewpoint. He/she could compare different profiles and know this person in more detail. Through these three profile image, he/she could find the suitable guide easily. Moreover, these three profiles from different viewpoints could be applied to other applications, such as a recommender system that helps to find a suitable match for something or someone.

Figure 4: A person's profile from personal viewpoint.

Figure 5: A person's profile from social viewpoint.

Figure 6: A person's profile from universal viewpoint.

Semantic Similarity Between Tags: Given two tags a and b, we could obtain two sets C_a and C_b including the neighboring relevant tags of a and b, respectively. The semantic similarity $Sim(a, b)$ between tag a and b is defined as Equation (10):

$$Sim(t_a, t_b) = \frac{|C_a \bigcap C_b|}{|C_a \bigcup C_b|} \qquad (10)$$

Co-occurrence We propose an co-occurrence measurement to compute the relationship between two different and irrelevant tags. For example, if an document is usually tagged as "Japan" and "Travel," we gain an insight to that user's preference and interest. If we never study any web technology, we cannot know the relationship among tags as

"ajax", "xml", and "javascript." These information cannot be known by commonsense semantic analysis, so we apply co-occurrence to calculate tag similarity. We consider two viewpoints when measuring tag similarity: (1) global co-occurrence and (2) personal co-occurrence. We apply global co-occurrence to measure common opinions and use personal co-occurrence to measure personal preference.

Global co-occurrence and personal co-occurrence are both measured by means of relative co-occurrence between tags, known as Jaccard coefficient. Let A and B be the sets of documents described by two tags, relative co-occurrence is defined as Equation (11):

$$RC(A,B) = \frac{|A \bigcap B|}{|A \bigcup B|} \tag{11}$$

where $RC(A,B)$ is the relative co-occurrence of A and B. $|A \bigcap B|$ is the number of document in which tags co-occur and $|A \bigcup B|$ is the number of resources in which appear in any one of the two tags. In other words, we compute the proportion of tag overlapping as tag similarity. The difference between global and personal are considered resources. Global co-occurrence calculate all tags given by all users in social bookmarking service and personal co-occurrence only considers tags given by a single user.

We use commonsense semantic analysis and co-occurrence measurement to calculate tag similarity. After analyzing similarity, we utilize graph cluster algorithm to cluster these tags. These clusters of tags would be useful for analysis and visualization.

Discussion

The interests or characteristics of people change over time. In addition to providing the alternative views, we should also consider temporal changes to model a person. A person's profile is a cumulative overview of interests, jobs, or skills. We need to track the pattern of transitions rather than capture one's profile at specific time slices. It will be interesting to observe that one's profile may dynamically change with friends, tendency, or other events in his/her life. Designing a visual presentation of such dynamic data is another challenge. A static picture is unable to represent temporal flow.

On the other hand, dynamic profile visualization is not enough. Not only do we need to visualize the user profile, but we must also automatically help the viewer find the right person. Depending on the purpose of a search, different match criteria should be considered. The matchmaker may define different *match functions*, such as *similar*, *complementary*, or *having interests overlap*, which should be further explored in our future work.

Conclusion

This paper presented an novel approach to user profiling based on the tags associated with one's social media. A profile should include not only personal (subjective) description about oneself, but also the opinions from one's social contacts as well as the global (objective) opinions of the general public. We defined the *tag capacity* for content as a measure of tag analysis over one's bookmark collection to determine the most descriptive tags. Furthermore, three alternative views, e.g. *personal*, *social*, and *global views*, are defined to offer a comprehensive profile of an individual. Through a *grouped tag cloud* to visualize a person's profile, viewers can easily focus on the most important and relevant information.

References

Collins, A., and Loftus, E. 1975. A spreading-activation theory of semantic processing. *Psychological Review* 82:407–428.

Furukawa, T.; Ishizuka, M.; Matsuo, Y.; Ohmukai, I.; and Uchiyama, K. 2007. Analyzing reading behavior by blog mining. In *AAAI*, 1353–1358.

Hasan-Montero, Y., and Herrero-Solana, V. 2006. Improving tag-clouds as a visual information retrieval interfaces. In *Proceedings of International Conference on Multidisciplinary Information Sciences and Technologies*.

Huang, T.-h., and Hsu, J. Y.-j. 2006. Beyond memories: Weaving photos into personal social networks. In Kaminka, G.; Pynadath, D.; and Geib, C., eds., *Modeling Others from Observations: Papers from the AAAI Workshop*, 29–36. Menlo Park, California: AAAI Press.

Kaser, O., and Lemire, D. 2007. Tag-cloud drawing: Algorithms for cloud visualization. In *Proc. WWW 2007 Workshop on Tagging and Metadata for Social Information Organization*.

Liu, H., and Maes, P. 2005. Interestmap: Harvesting social network profiles for recommendations. In *Proceedings of the Beyond Personalization 2005 Workshop*.

Liu, H., and Singh, P. 2004. Conceptnet: A practical commonsense reasoning toolkit. *BT Technology Journal*.