Personalization of Social Media

Maarten Clements
Faculty of Electrical Engineering, Mathematics and Computer Science
Delft University of Technology
The Netherlands
m.clements@tudelft.nl

Abstract: This article describes a framework that captures collaborative tagging systems, and derives from it an overview of user tasks that qualify for personalization in such a system. Major research areas have focused on some of these tasks, but we identify many more opportunities. We propose a collaborative model that combines collaborative filtering and information retrieval techniques in order to assists the user to achieve these tasks. Based only on the user's tags, this personalization model assumes that a user's tags identify this user's taste. Because many users do not only tag the content that matches their taste, we propose an evaluating experiment that shows if rating information can be used to adjust the users' taste profiles. This experiment is one of the steps to advance to a completely personalized model, integrating user preference, content annotations and people relations.

Keywords: Social Media, Personalization, Collaborative Tagging, Rating, Collaborative Filtering, Information Retrieval

1. INTRODUCTION TO SOCIAL MEDIA

Online social media have become respected tools for content sharing and relationship maintenance. People create digital identities to distribute videos or photos, share opinions about books and movies, or just maintain contact information of their friends. Multiple incentives exist for users who contribute to these social networks, it can however not be denied that the addition of social aspects in online databases has dramatically increased their popularity. The popularity of these collaborative systems has resulted in a substantial increase in user-generated content and user-generated metadata.

Collaborative social media exist in many shapes and designs. As a basic model for a collaborative system we use the flow chart shown in Figure 1. In this model, users can freely inject new content and assign an index for future retrieval. The indexed content can be retrieved by other people, using browse and search functionalities provided by the system. Any user who discovers content in this network can indicate if he approves with the index assigned to the content, or add his own index.

Indexing of content can occur in many different ways. Traditionally, large collections of content were only manageable using a consistent organization, often using hierarchical storage. With the introduction of ratings and tags in online databases, indexing of content has shifted from these strict hierarchies to subjective categorization. With ratings, users actively index certain content by the quality. Tagging allows users to assign keywords that they consider representative for the topic of the items. The staggering contribution of social network users makes that these new indexing methods result in practical database management tools.

This article considers online social media in which tags are used to index and retrieve the content and ratings can be assigned to indicate content preference. In many recently developed social networks that allow amateur content injection, the content itself contains very little information about its subject. The retrieval of home-made video fragments or pictures, completely depends on the indexing proficiency of the injector. To alleviate this problem, users should collaborate in the categorization of the content. We identify the difference between collaborative tagging systems (e.g. Del.icio.us[1], CiteULike[2]) and individual tagging systems (e.g. YouTube[3], Flickr[4]). In

[1] http://del.icio.us
[2] http://www.citeulike.org

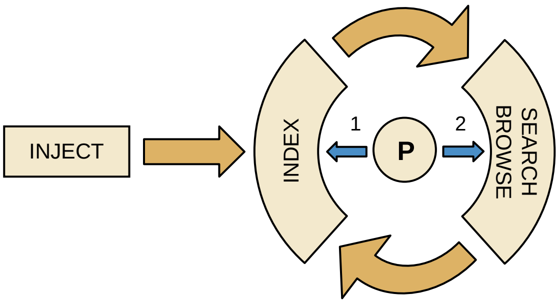

FIGURE 1: Collaborative Social Media. Content that is injected by any user can be retrieved and indexed by everyone. A personalization engine (P) can assist users in both the indexing (P1) and retrieval of content (P2).

collaborative tagging (CT), every user can tag any piece of content. Therefore, a tripartite relation exists that directly links user, item and tag. It has been shown that an individual tagging system is much harder to personalize, because items will have only a few tags assigned to them and users build up their preference profile much slower. Also, the problems that occur in tagging (Synonyms, Abbreviations, Ambiguity, Singular vs. Plural, Typos...) are much harder to deal with, if an item's tags are not aggregated over multiple users [2].

Figure 1 illustrates that personalization of collaborative tagging systems can assist users in both the indexing and retrieval phase. Most users are not trained to describe content using tags, and are insufficiently aware of tags in use by others. A personalized system suggests tags from the common vocabulary that fit the user's intention while remaining consistent with other users. As a result, users discover suitable tagging keywords more easily and, more important, inconsistent tagging behaviour is reduced. It has even been claimed that the suggestion of tags when a user is asked to label a certain item would lead to a more unanimously categorized database [2].

Navigation through tags provides an effective way to explore and discover content, while rating information improves the relevance ranking of the tagged items. To initiate the navigation, many current collaborative tagging systems make use of tag clouds, a visual representation, often based on the set of most popular tags. Popularity-based exploration is however limited, since different users may have very different preferences. Also, other ways to initiate browsing can be identified and different incentives exist besides the retrieval of content alone. It is worthwhile investigating how we can represent the user's preference using the tags and ratings he supplied, and how we can deploy this preference profile to personalize the existing tasks in a collaboratively indexed network.

2. SYSTEM PERSONALIZATION

Next to browsing possibilities and direct content recommendations, the social aspect of a content distribution system should be increased by displaying people that have shown interest in the same field as the user. It should allow users to click on other people, to explore their preferences and in this way get to know other network users. To meet these requirements, we suggest a framework that can always provide the user with relevant content, tags and people (see Figure 2). This framework gives an overview of the user tasks that qualify for personalization in a collaborative tagging system.

2.1 Data Structure

In a collaborative tagging system, the collection of tags can be visualized in a 3D matrix, where each position indicates if a user tagged an item with a specific tag (See Figure 3(a)). The matrix grows when people enter the network, content is introduced or someone uses a new tag. Because tagging data is usually very sparse, direct relations in the 3D matrix are often unreliable. We therefore sum over the 3 dimensions of the matrix to obtain:

- User-Tag (UT) matrix; indicating how many items each user tagged with which tag.
- Item-Tag (IT) matrix; indicating how many users tagged each item with which tag.
- User-Item (UI) matrix; indicating how many tags the users assigned to the items.

It cannot be assumed that the number of tags assigned to an item tells anything about the preference of the user towards that item. Therefore, the UI matrix should be replaced by rating data that directly represents the user's interest towards the items that he has tagged.

[3] http://www.youtube.com
[4] http://www.flickr.com

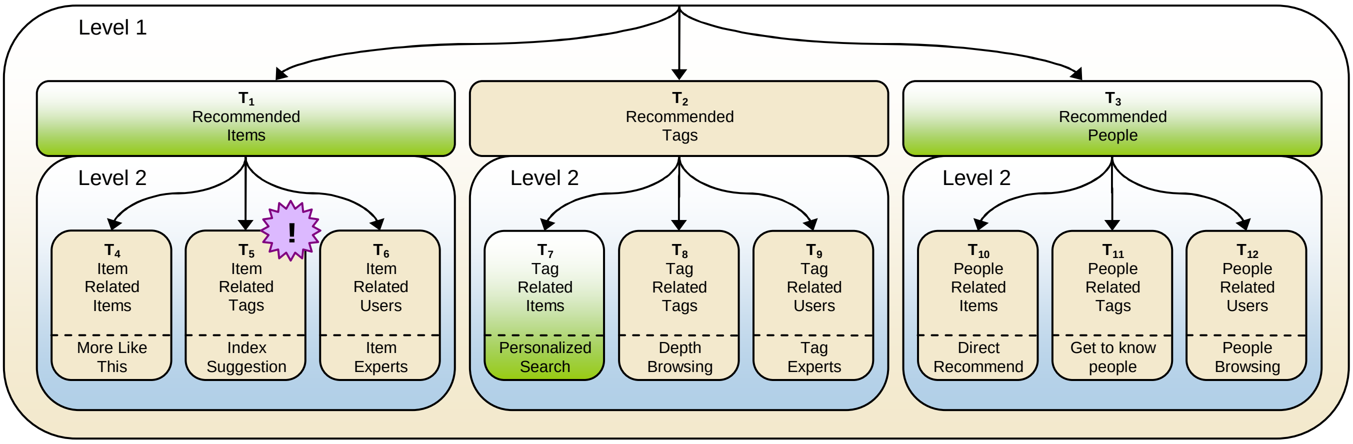

FIGURE 2: A schematic tree view of the tasks in a social browsing environment. Green tasks indicate major research fields until now. The task with a star (T_5) indicates the focus of this article. Level 1 shows the three tasks that apply to users that just enter the system (T_1-T_3). Many research activities have focused on the recommendation of items (T_1), often by first finding a group of similar users (T_3), which is known as Collaborative Filtering (1). Because most of these algorithms were developed on rated databases without tags, not much attention has gone to the recommendation of tags (T_2). Level 2 indicates the view on the network after the user has selected either an item, tag or user. In total, this model describes 12 tasks that apply for personalization in a collaboratively tagged social network. This includes common tasks like: recommendation of tags when interesting content has been found (T_5), retrieving relevant content by using tags as queries (T_7), getting help from experts on a certain topic (T_6,T_9), making new friends and using your friends to discover relevant content (T_{10}-T_{12}).

2.2 Personalization of Tasks

The twelve tasks from Figure 2 can be implemented using the three different projections of the UIT-matrix. By simply selecting profiles from the three matrices or computing correlations between these profiles, the view on the system can be adapted to the selected elements. If, for example, a piece of content has been selected, its tags can be shown by ranking the item's tag profile. This simple solution, however, ignores the information that we have about the user who clicked on the item. Many current systems ignore this information, and although these systems provide a high degree of browsing freedom, the represented content is often not based on the browsing behaviour of the person itself.

All tasks in the framework (Figure 2: T_1-T_{12}) can be personalized, by using the users own profile in the recommendation. To demonstrate the benefit of a personalized model we discuss *Personalized Indexing*; the suggestion of tags when a user wants to index interesting content (T_5). Looking at a popular system that currently provides assistance in the indexing phase, we see that Del.icio.us suggests both the tags that have been assigned to the item by other users (IT-profile) and the intersection of these tags with the target user's tags (UT-profile). The system provides a separate list for the suggested tags coming from the IT-profile and the tags coming from the UT-profile. Many users pick the most accurate tags from both these lists, so that the index matches both their personal categorization scheme and the popular opinion.

To aid the user in the selection of tags, a smooth combination of both information sources should be able to generate a single, more relevant list of tags. Based on the framework from Figure 2, we propose a *collaborative ranking model*, that serves as a basis for the personalization of all user tasks. This model computes a posterior probability, conditioned on the amount of information present at a certain level. When no user is logged on to the network, an item, tag or user ranking can only depend on the general popularity $p(\{u,i,t\})$ (the curly brackets indicate a selection of either u, i or t). If knowledge about the browsing user becomes available (*Level 1*), the ranking probability will be conditioned on this user's preferences: $p(\{u,i,t\}|u)$. By selecting an element from the presented interface (*Level 2*), the user provides information that can be used to improve the suggested ranking: $p(\{u,i,t\}|u,\{u,i,t\}_{L1})$. The model can easily be extended to deeper levels by conditioning the probability on more information. By adding an extra level below the tasks from *level 2*, queries can depend on combinations of multiple tags, items or users. Therefore allowing, for example, multiple keyword queries that are quite common in Information Retrieval.

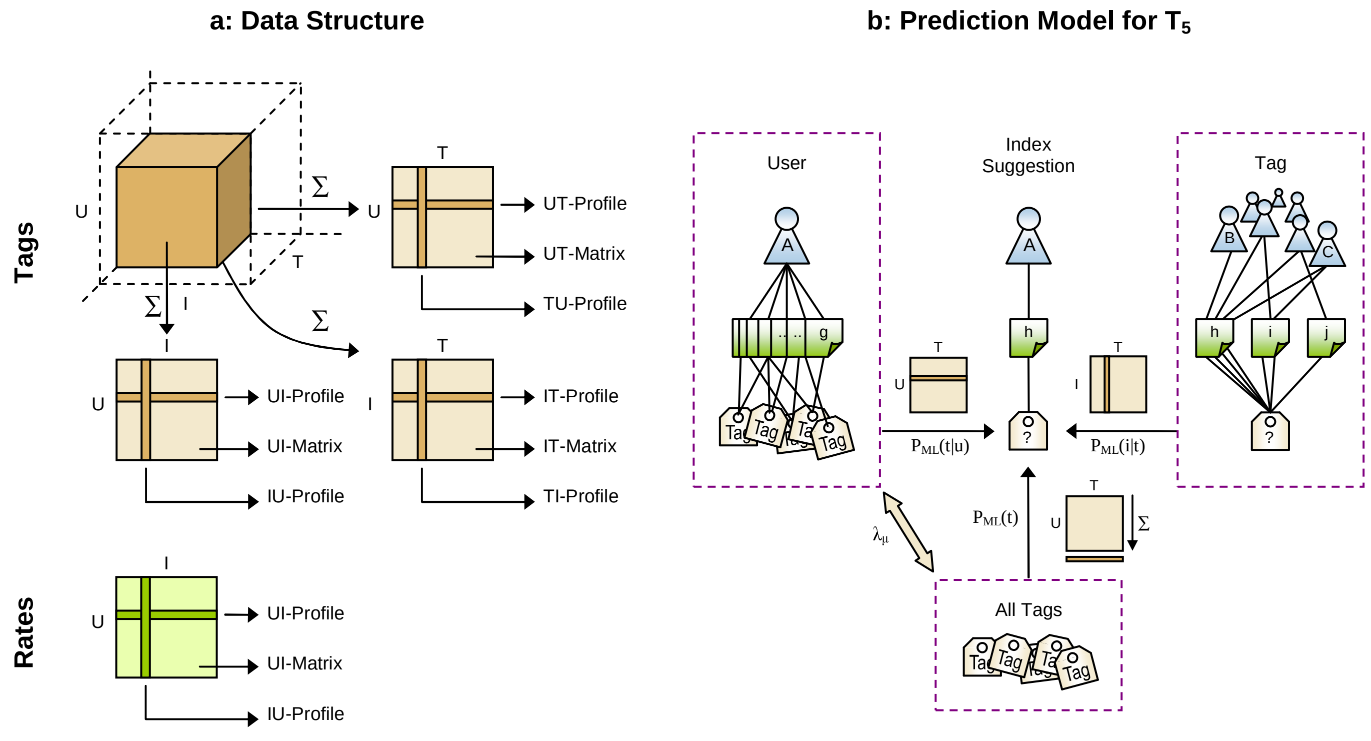

FIGURE 3: a) The User-Item-Tag matrix collects all tags that were assigned to the content in the system. The matrices used for tag-based personalization are obtained by summing the UIT-matrix in all dimensions. We refer to the resulting matrices as: UT-matrix, IT-matrix and UI-matrix. If we take a horizontal line from the UT-matrix we call this the UT-profile (read: "User's tag profile"). Summing over any of these matrices gives either the general tag popularity, item popularity or the network activity of a user. If rating information is present, it can be used to replace the tag-based UI-matrix. **b)** To suggest a relevant tag when a user (A) needs to index new found content (h), the model selects a tag from the UT-profile, that has often been assigned to the item before (considering all TI-profiles). To compensate for sparseness and cold start problems, the UT-profile is smoothed by the general tag popularity.

Applying our model to T_5, we suggest candidate tags t with high posterior probability $p(t|u,i)$, given that user u needs to label target item i. Bayes' rule can be used to rewrite this probability as:

$$p(t\,|\,u,i) = \frac{p(t\,|\,u)\,p(i\,|\,u,t)}{p(i\,|\,u)}$$

Using a language modeling approach with Bayesian smoothing to compensate for the sparse user profiles [4] we derive:

$$p(t\,|\,u,i) \propto \log(\lambda_u\, p_{ML}(t\,|\,u) + (1-\lambda)\, p_{ML}(t)) + \log p_{ML}(i\,|\,t), \quad \text{with}: \lambda_u = \left(\frac{\sum_t n(u,t)}{\mu + \sum_t n(u,t)} \right)$$

This ranking balances the general tag popularity $p_{ML}(t)$ against the personal preference $p_{ML}(t|u)$, while always taking into account the tags that have previously been assigned to the item $p_{ML}(i|t)$ (See Figure 3(b)). Initial experiments on tagging data from the del.icio.us network have shown that the optimal setting of the Bayesian smoothing parameter μ can improve the precision of the tag recommendation (on the top-3 ranked tags) by 7% over a completely personal ranking ($\mu = 0$) and by 9% over a ranking based on general tag popularity ($\mu \rightarrow \infty$).

3. PROPOSED EXPERIMENTS

Inspection of tagging data has shown that some people do not only tag the content they like, but are also willing to index content for the good of the community. Our proposed framework has made the assumption that the user's tag profile gives a correct representation of the user's preferences. This will result in bad recommendations for users who tagged content that does not interest them.

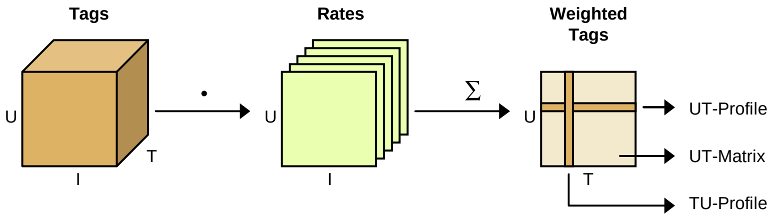

FIGURE 4: Weighing the tags by the rating supplied by the user gives a better representation of the user's preferences.

Many social networks allow users to indicate their preference toward specific content by giving the option to supply a rate. This rating information can be used to assign a weight to the tags that the user attached to this item, by multiplying the UIT-matrix with the users' rates before summing over the item dimension (see Fig. 4). In the ranking formula (Eq. 2) this would improve the effect of $p_{ML}(t|u)$. To evaluate if rating information can improve the representation of the user's taste in the tag profile we propose to repeat our experiments on a dataset from LibraryThing[5]. LibraryThing allows users to create an online catalog of the books they own or have read. A user can tag and rate all the books he adds to his personal library. The social aspects of this network give the user the opportunity to meet likeminded people and find new books that match their preference. The popularity of the system has resulted in a database, containing over 2 million unique works collaboratively added by more than 200.000 users. Currently we are aware of no other open network that contains this amount of collaborative tags and ratings.

We have collected a trace from the LibraryThing network containing 25.295 actively tagging users. As expected we see that the number of books in the users' catalogs follows a power law (see Fig. 5a). After pruning of this data set we retain 7564 users that have all supplied both rating and tags to at least 20 books. All books are annotated by at least 5 people resulting in 39.515 unique works. The user interface of LibraryThing allows users to assign ratings on the scale from a half to five. Half ratings can be given by clicking a star twice. The distribution in Figure 5b shows that half ratings are given about 4 times less than whole ratings. Figure 5c shows the relation between the rate and the number of tags given to an item. The upward trend shows that there is a slight correlation between these two variables. This graph also shows that books with half ratings tend to get more tags. This indicates that the half ratings are used by people who put more effort in the categorization of their books.

3.1. Future direction

The framework from Figure 2. has provided 12 different tasks that all need individual attention. The recommendation of content differs for example from the discussed task (T_5) in the fact that no content should be recommended that the user has seen before, whereas users prefer to use the same tags more than once. We need to work out the probabilistic model for the other tasks in the tree in order to get an adequate ranking method for each view on the network.

Other models can be applied to derive the user, item and tag probabilities, given a certain query. We will look at a random walk model in which the network nodes can be either a user, item or tag. The edges can be derived from the matrices in Figure 3a. The transition matrix $\boldsymbol{A}$ contains all 1-step probabilities, so that the probability of going from node j (at time t) to k (at time $t+1$), $P_{t+1|t}(k|j) = \boldsymbol{A}_{j,k}$. The initial state can now be represented as a vector $\boldsymbol{v}$ (with $\sum(\boldsymbol{v})=1$), in which the query elements can be assigned. By multiplying the state vector with the transition matrix, we can find the state probabilities after one step in the graph ($t+1$). Multi step probabilities can be found by repeating the multiplication $\boldsymbol{v}_{t+1} = \boldsymbol{v}_t \; x \; \boldsymbol{A}$ or using the n-step transition matrix $\boldsymbol{A}^n$.

The benefits of the random walk model include:

- Only one model is needed to execute all tasks from Figure 2. The final state vector only needs to be filtered to remove abundant results (e.g. already seen items).
- The system can easily be balanced between completely personal (trans. mat. $\boldsymbol{A}^1$) and completely popularity based (trans. mat. $\boldsymbol{A}^{inf}$).
- Different weights can be assigned to the query elements, e.g. in personalized tag recommendation T_5:$P(t|u,i)$, the personalization influence can easily be adapted by changing the proportions of u and i in the initial state vector. It can even be translated into the user interface where a user can dynamically change the weight of the query terms according to his needs.
- Simple computations using sparse matrix multiplication.

[5] http://www.librarything.com

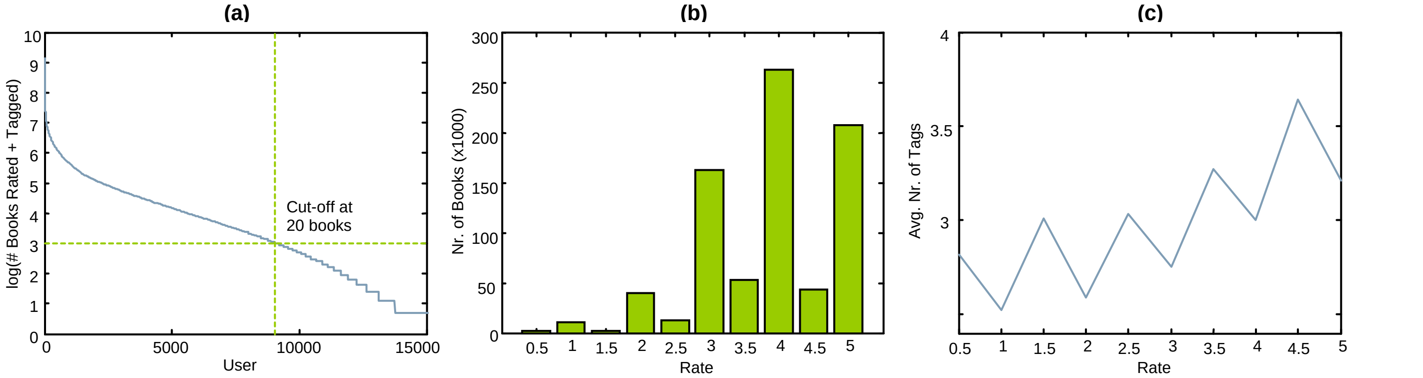

FIGURE 5: a) The number of rated and tagged books stored in the users' catalogs, sorted by size. **b)** The distribution of rating occurrences in the pruned data set. **c)** The average number of tags assigned, given the rating.

Design issues of the model:

- How to convert ratings to transition probabilities?
- How to balance the influence between ratings and tags?
- How to handle not rated/tagged items?

The random walk model allows easy integration of extra information. So far, the framework has ignored relations between users, while many social networks focus on this relation. Systems often allow users to actively assign 'friends' in the network, and many networks (like Ebay[6]) have implemented a 'trust' feature. In the random walk model, user relations can easily be added by allowing edges between users. Also, other sources like user demographics and item metadata could improve the estimation of transition probabilities.

Currently, we are collaborating with the development team of Tribler [3], a social P2P client that allows users to see other people's downloads and provides content recommendations. Using this client, we will build a dataset that maintains content ratings, tags and user friendships introduced by the network users. We will focus on the deployment of the social browsing model in a P2P environment. Currently, Tribler uses the similarity between the users' download profiles to create a social overlay that improves P2P download and enables social recommendations. The Tribler network can benefit from an improvement of the relevance ranking for either users, items or tags.

Acknowledgements
Many thanks to Jun Wang for the work on the probabilistic framework that I have built upon, Jie Yang for the del.icio.us dataset and work on the experiments, Arjen de Vries for fruitful discussions and mathematical insights, Johan Pouwelse for supplying the incentives and Marcel Reinders for the supervision and guidance of this work.

REFERENCES

[1] J. S. Breese, D. Heckerman, and C. Kadie (1998) Empirical analysis of predictive algorithms for collaborative filtering. *UAI-98: Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*, pp. 43–52, San Francisco, Morgan Kaufmann.

[2] S. A. Golder and B. A. Huberman (2006) The structure of collaborative tagging systems. Technical report, Information Dynamics Lab, HP Labs.

[3] J. Pouwelse, P. Garbacki, J. Wang, A. Bakker, J. Yang, A. Iosup, D. H. J. Epema, M. Reinders, M. van Steen, and H. Sips (2007) Tribler: A social-based peer-to-peer system. *Concurrency and Computation: Practice and Experience*, **19**:1–11.

[4] C. Zhai and J. Lafferty (2001) A study of smoothing methods for language models applied to ad hoc information retrieval. *SIGIR '01: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 334–342, New York, NY, USA. ACM Press.

[6] http://www.ebay.com