LikeMiner: A System for Mining the Power of 'Like' in Social Media Networks

Xin Jin
Dept. of Computer Science
University of Illinois at Urbana-Champaign
201 N. Goodwin Ave.
Urbana, IL USA
xinjin3@illinois.edu

Chi Wang
Dept. of Computer Science
University of Illinois at Urbana-Champaign
201 N. Goodwin Ave.
Urbana, IL USA
chiwang1@illinois.edu

Jiebo Luo
Kodak Research Laboratories
Eastman Kodak Company
1999 Lake Avenue
Rochester, NY USA
jiebo.luo@kodak.com

Xiao Yu
Dept. of Computer Science
University of Illinois at Urbana-Champaign
201 N. Goodwin Ave.
Urbana, IL USA
xiaoyu1@illinois.edu

Jiawei Han
Dept. of Computer Science
University of Illinois at Urbana-Champaign
201 N. Goodwin Ave.
Urbana, IL USA
hanj@cs.uiuc.edu

ABSTRACT

Social media is becoming increasingly ubiquitous and popular on the Internet. Due to the huge popularity of social media websites, such as Facebook, Twitter, YouTube and Flickr, many companies or public figures are now active in maintaining pages on those websites to interact with online users, attracting a large number of fans/followers by posting interesting objects, e.g., (product) photos/videos and text messages. '**Like**' has now become a very popular *social* function by allowing users to express their like of certain objects. It provides an accurate way of estimating user interests and an effective way of sharing/promoting information in social media. In this demo, we propose a system called LikeMiner to mine the power of 'like' in social media networks. We introduce a heterogeneous network model for social media with 'likes', and propose 'like' mining algorithms to estimate representativeness and influence of objects. The implemented prototype system demonstrates the effectiveness of the proposed approach using the large scale Facebook data.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database applications –Data Mining; H.3.4 [**Information Storage and Retrieval**]: Systems and Software –Information networks; J.4 [**Computer Applications**]: Social and Behavioral Sciences

General Terms

Algorithms, Design, Experimentation, Human Factors

Keywords

Data mining, social media, information network, like, influence analysis, recommendation, ranking

KDD'11, August 21–24, 2011, San Diego, California, USA.

1. INTRODUCTION

Social media has become one of the most popular web and mobile applications. According to a study[1] from the Nielsen Company, in 2010 the world spent over 110 billion minutes, which equals to 22 percent of all time online, on social-media related websites, such as Facebook, Twitter, YouTube and Flickr. The time spent on these sites by an average visitor increased by 66% over a year ago. Because of the huge popularity and rich personal information available on social media websites, companies or public figures have now being increasingly willing and active in maintaining pages on those websites to interact with online users, attracting a large number of fans/followers by posting interesting (product) photos/videos or text messages.

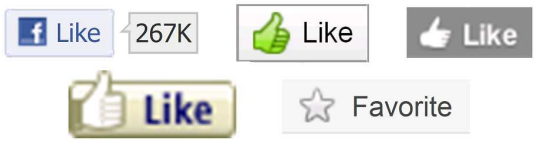

Figure 1: The like/favorite button for Facebook, YouTube, theAtlantic, Amazon and Flickr, respectively.

'**Like**' has recently become a very popular *social* function on the Internet. Many social media websites, such as Facebook, Twitter, YouTube and Flickr, provide the like/favorite

[1]Nielsen. http://blog.nielsen.com/nielsenwire/global/social-media-accounts-for-22-percent-of-time-online Accessed 2011

button to allow users to express their like of the *objects* (such as text messages, comments, webpages, photos and videos) posted either by personal users or companies and public figures. Many traditional websites now also include the 'like' button, either as a plugin from social media websites or created by themselves, to help promote the webpages. Figure 1 shows the like/favorite buttons used by Facebook, YouTube, theAtlantic, Amazon and Flickr.

If a user clicks 'like' associated with an object, this directly indicates that s/he is highly interested in the object. So the 'like' function provides a more accurate way of estimating user interests than non-direct indicators, such as user-service interaction [20]. Additionally, in some social networks, e.g., Facebook, when a user clicks 'like' to an object, such action will be immediately shared to his/her friends (under allowed privacy setting). So the 'like' function also provides a useful and effective way of sharing or promoting information in social media. Actually, sharing by 'like' may have higher influence because people may pay more attention to the objects *liked* than simply *shared* by friends.

In this demo, we propose a system called LikeMiner to mine the power of 'like' in social media networks. The major contributions are: (a) we introduce a heterogeneous like network model for social media; (b) we propose to construct both visual and textual topic space for social media objects; (c) we propose mining algorithms to estimate representativeness and influence in the social media; and (d) we implement a working system to demonstrate the effectiveness of the proposed approach.

2. 'LIKE' NETWORK MODEL

We model a social media network with 'like' function as a heterogeneous information network $G = \langle V, E \rangle$. V is the set of nodes from different types, such as users (U), pages (P), and posted objects (O) (such as text messages, comments, photos/videos, accompanied by the posting time). E is the set of directed edges between nodes, including friendship/following links between users, fan links between users and pages, the posting links between users/pages and objects, the like/favorite links between users and the posted objects. Photos are indirectly-linked together by content similarity (dashed lines).

Figure 2 shows an example of such network within Facebook. A user can post an object to his/her wall, or to any company/public-figure pages s/he is a fan. Such object can be liked by his/her friends or non-friends who are also fans of that page. A page can also post its own objects and those interesting ones will be liked by its fans.

3. SYSTEM FRAMEWORK

As shown in Figure 3, the system architecture works as follows. (1) Forming a network model for social media with likes; (2) extracting the topic distribution for the objects, both visual and textual; (3) performing link mining based on the network structure and the object topic distribution; and (4) providing the web interface for browsing, keyword-based topic query and interest-based recommendation, such as recommending product photos or company pages to users who may potentially like them.

3.1 Object Topic Extraction

A *topic space* is a d-dimensional real number space $\mathbb{R}^d$

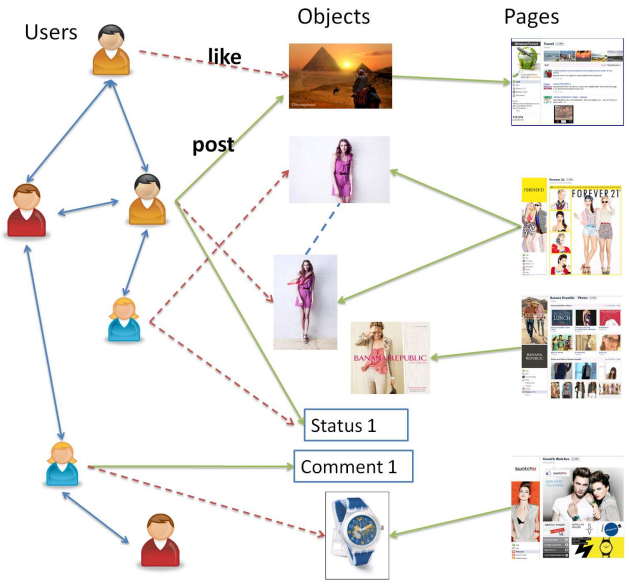

Figure 2: A heterogeneous network model for social media with 'like' function, using Facebook as an example. A blue bidirectional arrow is the friendship link. A red dashed arrow is a *like* action, while a green arrow denotes a *post* action, annotated with the time stamp when it was posted. The dashed blue line shows that the two photos are visually similar.

that encodes some cognitive classification systems such as ontology. Each dimension represents a topic class in the system. For example, in a human-edited Web directory [1] , each category in the directory can be used as a topic class.

A *topic vector* is a vector in the topic space defined above that represents the cognitive property of an object. The component in each dimension indicates how strong the object is associated with the corresponding topic. For example, (Movie:0.5, Soccer:0.3) is a short representation of a sparse topic vector where all the weights are zero except for two topics.

We extract visual topics space for the photos/videos based on the content features, and textual topics space for the textual objects, such as comments and Facebook wall statuses.

3.1.1 *Visual Topic Extraction*

We can extract different types of image content features [9] [5] [6], such as color histogram, color correlogram [7], texture features [2] [13] [14] [16] [17], Gabor features [15] [18], edge histogram [12] and SIFT [11], to represent the photos from different perspectives. Each type of feature is in a high-dimension space, with the number of dimensions ranging from tens (e.g., color histogram) to hundreds (e.g., SIFT).

Based on the extracted features, we apply our recent algorithm GAD [8] to perform large scale clustering of the photos for each feature space, and treat each cluster as a visual topic. The topic distribution of a photo is based on the similarity of the photo to each cluster center.

3.1.2 *Textual Topic Extraction*

The task of textual topic extraction is to transform the

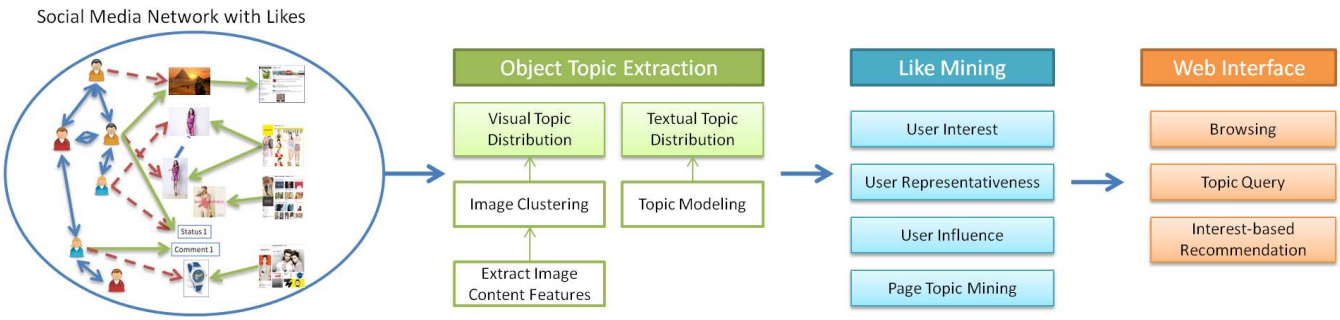

Figure 3: LikeMiner System Architecture.

text associated with a social media object into the most salient topics that occur in the text. A variety of text modeling approaches are possible here, including categorization approaches that classify text into a pre-defined ontology of topics, and topic modeling approaches that automatically find semantically meaningful clusters [3]. For the categorization-based approach, we chose the publicly available DMOZ/ODP hierarchy [1] as our ontology, because of its comprehensive coverage and the availability of labeled webpage data at each hierarchical node.

3.2 User Interest Mining

In this system we use the topic of the objects that a user posted or liked to infer the interest of that user. For every user u, his/her interest is modeled as a linear combination of the topic vector from those objects.

The user topic vector $\mathbf{t}(u)$ may contain some noisy interests. Also, the absolute value in each dimension is highly affected by the number of objects a user liked or posted. We can use smoothing and normalization techniques to transform the user topic vector to accommodate different applications.

3.3 User Representativeness and Influence Mining

Given a specific topic, we mine two kinds of knowledge about users. The first kind is the users who are representative of the opinions among friends in this topic, i.e., the objects s/he likes are also liked by many other users. The second is the influence between users. Some user has strong influence on his/her friends on this specific topic, i.e., the comments, pictures and status about this topic s/he posted are liked by friends.

We mine such knowledge in the following steps. First, we filter the network so that we only keep the users with interest on the given topic and the objects of the corresponding topic. Then, we construct two directional user graphs based on a user-object network and a user-user social network. One is based on co-like relationship, and the other is based on like-post relationship.

We use the co-like graph to mine the representativeness with a PageRank algorithm. We define the random walk probability from one user u to another v according to the co-liked objects and their topics.

The like-post graph is used to perform influence analysis similar to Tang et al. [19]. The difference is that we use directional links based on like-post relationship and we define the weight on every edge from u to v based on the posted and liked objects.

We can further use the result to find a small set of influential users on the network on the given topic based on influence maximization[10]. Scalable algorithms have been developed to solve that problem given the network diffusion model [4]. These algorithms require the network structure and the influence probability as input. We use the influence analysis results r_{uv} to feed the MIA algorithm in [4], and generate the final output of seed users.

3.4 Page Topic and Representativeness Mining

To determine the topic of a page, we aggregate the topic vector from both the objects contained in the page and the users who like the page. Although the topic of one object or one user can be inaccurate, when we collect them together and find the majority, we expect to remedy the sparsity and noises with the help of collective intelligence.

We also mine the topic-specific representativeness of pages in the sense that the more a page has common fans with other pages on the same topic, the more likely this page is a good representative of that topic.

4. DEMONSTRATION

We showcase the LikeMiner prototype system using Facebook as the example application. In addition to 500 million users, there are over 14 million (the number is keep growing) Facebook pages from various categories, such as company, product/service, musician/band, politician, artist, athlete and movie. There are on average 2 million fans for the top 3000 popular pages. The most popular one (Texas Hold'em Poker) has over 40 million fans. Fans not only can see posts submitted by the page, but also can like or post comments/photos/videos to the page. Most Facebook pages are public accessible, we integrate them together to construct a large scale heterogeneous social media network.

Figure 4 shows the web interface of the LikeMiner system. The website is supported by Apache and MySQL. The user interface is supported by PHP and HTML. In this figure, the user submits a topic query "clothing", the system returns the top users (with the user scope option of only friends, friends of friends or everyone) based on interest relevance, representativeness or influence, respectively. The system will also return top relevant photos and pages based on the number of likes, relevance score or representativeness.

In addition, we can calculate the similarity based on the

user topic interest distribution and photo/page topic distribution to recommend photos or pages to the user, which can be used for accurate advertising targeting on social media websites.

Figure 4: The LikeMiner prototype system. The user profile faces are mosaicked to protect privacy.

Note that this demo uses Facebook as the example since it is currently the most popular social media website. However, the technique proposed can be easily applied/extended to other social media websites, such as Flickr and YouTube. For YouTube, we need to extract the video features to help build the visual topic space.

5. ACKNOWLEDGMENTS

This work was sponsored by Eastman Kodak Company and supported in part by NSF grant IIS-09-05215, MURI award FA9550-08-1-0265 and NS-CTA W911NF-09-2-0053. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing official policies, either expressed or implied, of the sponsors.

6. REFERENCES

[1] *The Open Directory Project (ODP),* `http://www.dmoz.org`.

[2] S. Aksoy and R. M. Haralick. Textural features for image database retrieval. In *CBAIVL '98: Proceedings of the IEEE Workshop on Content - Based Access of Image and Video Libraries*, page 45, 1998.

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[4] W. Chen, C. Wang, and Y. Wang. Scalable influence maximization for prevalent viral marketing in large-scale social networks. In *KDD'10*, Washington D.C., July 2010.

[5] R. Datta, D. Joshi, J. Li, and J. Z. Wang. Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Surveys*, 40(2):1–60, April 2008.

[6] T. Deselaers, D. Keysers, and H. Ney. Features for image retrieval: an experimental comparison. *Information Retrieval*, 11(2):77–107, 2008.

[7] J. Huang, S. R. Kumar, M. Mitra, W.-J. Zhu, and R. Zabih. Image indexing using color correlograms. In *Proceedings of the 1997 Conference on Computer Vision and Pattern Recognition (CVPR '97)*, page 762, 1997.

[8] X. Jin, S. Kim, J. Han, L. Kao, and Z. Yin. A general framework for efficient clustering of large datasets based on activity detection. *Statistical Analysis and Data Mining*, 4(1):11–29, 2011.

[9] X. Jin, J. Luo, J. Yu, G. Wang, D. Joshi, and J. Han. iRIN: image retrieval in image-rich information networks. In M. Rappa, P. Jones, J. Freire, and S. Chakrabarti, editors, *WWW*, pages 1261–1264. ACM, 2010.

[10] D. Kempe, J. Kleinberg, and E. Tardos. Maximizing the spread of influence through a social network. In *KDD'03*, pages 137–146, 2003.

[11] D. G. Lowe. Object recognition from local scale-invariant features. In *Proceedings of the International Conference on Computer Vision-Volume 2*, page 1150, 1999.

[12] R. C. Ltd. MINDS's descriptors for still images - spatial edge distribution descriptor. *ISO/IEC/JTC1/SC29/WG11, Lancaster, UK, Feb*, page 109, 1999.

[13] K. Muller and J. R. Ohm. Wavelet-based contour descriptor. *ISO/IEC/JTC1/SC29/WG11, Lancaster, UK, Feb.*, page 567, 1999.

[14] A. V. Nevel. Texture synthesis via matching first and second order statistics of a wavelet frame decomposition. *Image Processing, International Conference on*, 1:72, 1998.

[15] M. Park, J. S. Jin, and L. S. Wilson. Fast content-based image retrieval using quasi-gabor filter and reduction of image feature dimension. In *SSIAI '02: Proceedings of the Fifth IEEE Southwest Symposium on Image Analysis and Interpretation*, page 178, 2002.

[16] A. R. Rao and G. L. Lohse. Towards a texture naming system: identifying relevant dimensions of texture. In *VIS '93: Proceedings of the 4th conference on Visualization '93*, pages 220–227, Washington, DC, USA, 1993. IEEE Computer Society.

[17] Y. M. Ro, S. Y. Kim, K. W. You, M. Kim, and J. Kim. Texture description using atoms of matching pursuits. *ISO/IEC JTC1/SC29/WG11, Lancaster, UK, Feb*, page 612, 1999.

[18] D. M. Squire, W. Muler, H. Muler, and T. Pun. Content-based query of image databases: inspirations from text retrieval. *Pattern Recognition Letters*, 21(13-14):1193–1198, 2000. Selected Papers from The 11th Scandinavian Conference on Image.

[19] J. Tang, J. Sun, C. Wang, and Z. Yang. Social influence analysis in large-scale networks. In *KDD '09*, 2009.

[20] S.-H. Yang, B. Long, A. Smola, N. Sadagopan, Z. Zheng, and H. Zha. Like like alike: joint friendship and interest propagation in social networks. In *Proceedings of the 20th international conference on World wide web*, WWW '11, pages 537–546, 2011.