Building the Search Pattern of Social Media User Based on Cyber Individual Model

Zheng Xu, Xiao Wei, Dongmin Chen, Haiyan Chen and Fangfang Liu

Abstract As the Web enters Big Data age, users and search engines may find it more and more difficult to effectively use and manage such big data. On one hand, people expect to get more accurate information with less search steps. On the other hand, search engines are expected to incur fewer resources of computing, storage and network, while serving the users more effectively. After more and more personal data becomes available, the basic issue is how to generate Cyber-I's initial models and make the models growable. The ultimate goal is for the growing models to successively approach to or become more similar as individual's actual characteristics along with increasing personal data from various sources covering different aspects. In this paper, we propose the concept of search pattern, summarize search engines into three search patterns and compare them in order to seek the more efficient one. We propose a new search pattern termed as ExNa, which can be incorporated into search engines to support more efficient search with better results.

Keywords Search pattern · Social media · Cyber individual model

Z. Xu (✉)
Tsinghua University, Beijing, China
e-mail: xuzheng@shu.edu.cn

Z. Xu
The Third Research Institute of Ministry of Public Security, Shanghai, China

X. Wei
Shanghai Institute of Technology, Shanghai, China

D. Chen
Software College of Northeastern University, Shenyang, China

F. Liu
Shanghai University, Shanghai, China

H. Chen
East China University of Political Science and Law, Shanghai, China

J.C. Hung et al. (eds.), *Frontier Computing*, Lecture Notes
in Electrical Engineering 375, DOI 10.1007/978-981-10-0539-8_3

1 Introduction

As the Web enters Big Data age, users and search engines may find it more and more difficult to effectively use and manage such big data. On one hand, people expect to get more accurate information with less search steps. On the other hand, search engines are expected to incur fewer resources of computing, storage and network, while serving the users more effectively. New types of search engines are emerging to solve the problem. In particular, faceted search [1, 2], ontology-based search [3], concept-based search [4], and rule-based search [5] all aim to improve search engines in some aspects and contribute to the development of so-called “next generation” search engines (NGSEs).

With rapid advances of computing and communication technologies, we are stepping into a completely new cyber-physical integrated hyper world with digital explosions of data, connectivity, services and intelligence. As individuals facing so many services in the digitally explosive world, we may not be aware of what are the most necessary or suitable things [6–9]. Hence, the appearance of Cyber-I, short for Cyber-Individual, is a counterpart of a real individual (Real-I) to digitally clone every person [10, 11]. The study on Cyber-I is an effort to re-examine and analyze human essence in the cyber-physical integrated world in order to assist the individuals in dealing with the service explosions for having an enjoyable life in the emerging hyper world.

After more and more personal data becomes available, the basic issue is how to generate Cyber-I’s initial models and make the models growable. The ultimate goal is for the growing models to successively approach to or become more similar as individual’s actual characteristics along with increasing personal data from various sources covering different aspects. The focus of this research is on the initialization and growth of Cyber-I’s models. The initial models are generated based on the personal data acquired in a Cyber-I’s birth stage, while the growing models are built with the personal data continuously collected after the birth. We proposed three mechanisms for Cyber-I modeling to enable the models growing bigger, higher and closer successively to its Real-I.

A big question here is then that in order to achieve NGSEs, what types of search patterns should NGSEs support. With an aim to help find a possible answer to this big question, in this paper we adopt an inside-out approach by first defining *Search Pattern* (*SP*) as the combination of index structure, user profiles, and interaction mechanism, which can describe the features related to the search process more comprehensively, including those of NGSEs. Then, we summarize current search engines into three types of search patterns. By comparing and analyzing different patterns, we try to identify what features a “next generation” search engine (NGSE) should have and what search patterns NGSEs should support. Based on this, we propose a new search pattern named ExNa by defining its model and basic operations. To validate the newly proposed ExNa search pattern, we conduct experimental studies upon a semantic search engine named NEWSEARCH, and the results show that KNOWLE equipped with ExNa can improve the holistic efficiency of the search system. A search pattern may be good at a special aspect of a search engine, such as the precision of searching, the storage of index, the I/O, and so on. ExNa is good at the holistic efficiency when compared with search engines of other search patterns.

In this paper, we propose the concept of search pattern, summarize search engines into three search patterns and compare them in order to seek the more efficient one. We propose a new search pattern termed as ExNa, which can be incorporated into search engines to support more efficient search with better results.

2 Related Work

User models are also known as user profiles, personas or archetypes. They can be used by designers and developers for personalization purposes so as to increase the usability and accessibility of products and services. With the development of personalized systems, like e-learning systems, a lot of personal data can be collected. In order to find some personal features to give appropriate advices or recommendations, the user model should be established in service systems. However, many such kind of user models is application-specific or service-specific which cannot be used by other applications/services. To overcome this barrier of the user models between different applications, a generic user model system (GUMS) was proposed to support interoperability among different user modeling systems [12]. The GUMS is able to exchange contents of user models, and use the exchanged user's information to enrich the user experience. Life logging is utilized to automatically record user's life events in digital format. With continuously capturing contextual information from a user and the user's environment, personal data increases fast and becomes huge. The most of lifelog systems are putting more emphases on personal data collection, storage and management [13]. Lifelong user modeling is trying to provide users such models accompanied with users' whole life [14]. This idea or vision is attractive, but no general mechanism has been made and no practical system has been built yet. Lifelong machine learning (LML), received great attention in recent years, is to enable an algorithm or a system to learn tasks from more domains over its lifetime [15].

3 The Search Pattern

ExNa is not a simple integration of the Narrow SP and the Expand SP. ExNa is expected to have a free styled interaction, a more efficient index structure which should be rich semantics, less storage, and abundant interaction paths, and a flexible user profile to support all kinds of service. Some conflicts should be resolved in ExNa, such as the conflict between the rich semantics and the huge storage.

And some problems should be solved in ExNa too, such as how to realize the free styled interaction, and how to build a flexible user profile to support all kinds of services.

Although ExNa is a little like the integration of Narrow SP and Expand SP, it just includes the interaction paths of Narrow SP and Expand SP. As the definition of Search Pattern shown, SP consists of three parts and the search path is only the representation of the entire SP.

Based on the discussions of Linear SP, Narrow SP and Expand SP, we compares them as per the structure of index, the storage of index, the semantics of index, the interactive mechanism and user profiles. We strive to find a new search pattern by integrating the advantages of the current SPs as many as possible. Clearly, Narrow SP and Expand SP work in vertical and horizontal directions, respectively. Narrow SP may rapidly narrow the search scope with the support of hierarchical index structure. Expand SP may expand the search based on some semantic relations, thereby facilitating user search with fuzzy terms. Take both vertical and horizontal directions into account, the index of the new SP should be a structure of multi-layered, in which different layers denote the indices of different granularities. Besides, the web resources of the same layer should be organized as a semantic link network. We name the index structure as the multi-layered semantic link network index structure. Rich semantics should be included in the new index structure to support efficient search. The semantic relations between layers support the narrow search. The semantic link network may hold several kinds of semantic relations to support the expand search. Storing rich semantic information needs more storage than the inverted index. The multi-layered and community structure in semantic link network may reduce the storage of index to a large extent. We expect the storage space to be at the medium level which is much less than a single layer network structure such as Expand SP. With the support of the multi-layered semantic link network index structure, a user may interact with the index from both vertical and horizontal directions, which form a free-styled interaction mechanism. To support the free-styled interaction, the proper structure of user profiles should be a multi-layered network too, so as to record user.

4 The Basic Data for Social Media Profile

SNS profile. The online social networking service (SNS), like Facebook.com, is a great way to find out more about you, which allows anyone with an email address to create a profile complete with pictures and a variety of specific personal information. Personal information is voluntarily supplied by the user and usually contains information such as Major, Hometown, Relationship, Status, Political Views, Interests, Favorite Music/Movies/Books/Quotes, and an “About Me” section which contains a short description of the user someone you have just met. The SNS profile play an important role during the initialization of Cyber-I modeling since it contains some context information that is able to be utilized. For instance, taking a user’s age, occupation or hometown into consideration will better locate the user or give the user a better service or more applications will be added in order to generate more personal data.

Preference Choice. It has long period study/research in the area of psychology, and psychological research give us proof that the recognition of user preferences could reflect something deep inside the user, such as the characteristics, the trait. And such preference could also lead to influence the selection and instantiation of the action that achieve the user's target. In this thesis, we start from the simple color preference, which may not be sensitive for someone's privacy concern and generally speaking, everyone has his/her own loved color. Color preference is an important aspect of visual experience that influences a wide spectrum of human behaviors. Secondly, we suggest the user to choose the other optional preference choices, which are available as Foods, Sport, Movie and Music. If users are willing to choose those (we are strongly suggest to do this), the model can get and know your properties of different aspects in order to generate a better initial model for you and provide you more services/apps. The function of preference choices will be talked in detail in the next section.

Browsing History, App Usage and Activity Tag. In order to fetch the information concerning the user's activity on PC, we make use of the software "Manic Time" to implement those functions, which could generate the data into the different CSV files. The files can be uploaded manually into the database and can be processed by our Java program in processor database. We calculate the total time and the times you open one software during your working on PC. Meanwhile, the frequently visited website can also be analyzed through this Java program. After analyzing the CSV files, the consequences can be demonstrated on the form of pie chart or bar graph based on the Google Chart Visualization API. What's more, we can generate further results, such as the top 3 favorite websites, what application or even what kinds of information are preferred. For the professional like employees and students who are in front of computer every day, the activity tags like "go for lunch" "afternoon nap", "time for dinner" can also be demonstrated in the results and able to be stored into database for modeling.

Movement log. In order to collect the movement log of the number of steps of the day, UP of Jawbone Company, which is a wearable activity recording device, can be used. Further, it is possible to synchronize and visualize the data at any time measured by using the UP smartphone application. In addition, since it measurably every day, logging your exercise conditions on an ongoing basis. UP is possible to use about 1 week on a single charge which is designed that user can wear everyday with waterproof function and with just 22 g weight body in wristband. In this research, the number of steps was collected and through synchronizing with smartphone to transfer data to the server of JAWBONE, steps and exercise situation and the consumption of calories each date can be stored. Further, it is possible to access the home page of JAWBONE, obtaining CSV format data in the account page when have been registered. Analysis is performed to get the number of steps for knowing the motion state through the data obtained from the UP in our present study.

5 Conclusions

To overcome the shortcomings of traditional search engines, many new search engines are designed based on some new technologies. Each search engine has its advantages and disadvantages. So we try to summarize the current search engines to find the features of search engines of next generation. In this paper, we propose the concept "Search Pattern" to describe the most important features of search engines. We classify current search engines into three Search Patterns: Linear Search Pattern, Narrow Search Pattern, and Expand Search Pattern. We present a novel search pattern ExNa based on the comparison of Linear Search Pattern, Narrow Search Pattern, and Expand Search Pattern. Then, we model ExNa and definition its basic operations to help developing search engines of next generation.

Acknowledgments This work was supported in part by the National Science and Technology Major Project under Grant 2013ZX01033002-003, in part by the National High Technology Research and Development Program of China (863 Program) under Grant 2013AA014601, 2013AA014603, in part by National Key Technology Support Program under Grant 2012BAH07B01, in part by the National Science Foundation of China under Grant 61300202, 61300028, in part by the Project of the Ministry of Public Security under Grant 2014JSYJB009, in part by the China Postdoctoral Science Foundation under Grant 2014M560085, the project of Shanghai Municipal Commission of Economy and Information under Grant 12GA-19, and in part by the Science Foundation of Shanghai under Grant 13ZR1452900, 12ZR1411000.

References

1. ICSTI Insight: Next generation search. http://www.icsti.org/IMG/pdf/insight_2010_july.pdf
2. Tunkelang D (2009) Faceted search. In: Synthesis lectures on information concepts, retrieval, and services, vol. 1(1), pp 1–80
3. Alisi T, Bertini M, D'Amico G, Del Bimbo A, Ferracani A, Pernici F, Serra, G (2009) Sirio: an ontology-based web search engine for videos. In: Proceedings of the 17th ACM international conference on multimedia, pp 967–968
4. Shehata S, Karray F, Kamel M (2007) Enhancing search engine quality using concept-based text retrieval. In: Proceedings of IEEE/WIC/ACM international conference on web intelligence, pp 26–32
5. Pilz T, Luther W, Ammon U, Fuhr N (2006) Rule-based search in text databases with nonstandard orthographym. Literary Linguist Comput 21(2):179–186
6. Wang L, Ke L, Liu P (2015) Compressed sensing of a remote sensing image based on the priors of the reference image. IEEE Geosci Remote Sens Lett 12(4):736–740
7. Liu P, Yuan T, Ma Y, Wang L, Liu D, Yue S, Kolodziej J (2014) Parallel processing of massive remote sensing images in a GPU archi-tecture. Comput Inform 33(1):197–217
8. Wang L, Ke L, Liu P, Ranjan R, Chen L (2014) IK-SVD: dictionary learning for spatial big data via incremental atom update. Comput Sci Eng 16(4):41–52
9. Wang L, Geng H, Liu P, Ke L, Kolodziej J, Ranjan R, Zomaya AY (2015) Particle swarm optimization based dictionary learning for remote sensing big data. Knowl-Based Syst 79:43–50
10. Yen NY, Ma J, Huang R, Jin Q, Shih TK (2010) Shift to Cyber-I: reexamining personalized pervasive learning. In: Proceedings of the 3rd IEEE/ACM International conference on cyber. Physical and social computing, pp 685–690 December 2010

11. Wei J, Huang B, Ma J (2009) Cyber-I: vision of the individual's counterpart on cyberspace. In: Proceedings of the IEEE international conference on dependable. Autonomic and secure computing, pp 295–302
12. Levene M (2011) An introduction to search engines and web navigation. Wiley
13. Hu WC, Chen Y, Schmalz MS, Ritter GX (2001). An overview of the world wide web search technologies. In: Proceedings of the 5th world multi-conference on system, cybernetics and information, pp 22–25
14. Cutting D, Pedersen J (1989) Optimization for dynamic inverted index maintenance. In: Proceedings of the 13th annual international ACM SIGIR conference on research and development in information retrieval, pp 405–411
15. Jain AK, Murty MN, Flynn PJ (1999) Data clustering: a review. ACM Comput Surv 31 (3):264–323