THE MODES OF DATA DEVELOPMENT IN THE INTERNET AGE

Yuxian Wu

Institute of Economics CASS, Beijing, 100836
Email: wuyx@cass.org.cn

ABSTRACT

It is historical that data development has its own mode (collect, treatment, delivery, store, and use), from Manual mode, Mechanism mode, and Electronic mode, now to the Network mode. And search engine plus self-learning is the advanced mode of data development. Network mode has also been changing, the underlying motivation exits in the development & progress of Internet itself. There are two huge trends force the mode of data development to face new challenge & make decision. One is the revolution resulted by the change of the user market need & represented by Web2.0. Another is the revolution resulted by technological developing tendency & represented by Grid. Squeezed by the two huge trends from opposite directions, the lagging, crude and inefficient mode will change revolutionarily forced by wise decision-making or silent market. As for data-development mode, the change of technology & operation need the change of game rule simultaneously. So eliminating barriers, promoting resource-sharing, rationalize relations of market/non-market is to be a big inescapable work

Keywords: Data, Development-mode, Internet, Grid, Data mining, Web2.0

1 INTRODUCTION

It is well known that the information era is marked by computer technology as its originating feature and Internet technologies for its present climax. Yet the mode of data/information development (collection, processing, transmission, storage, use, etc.) has a very ancient history. From recording by tying a string, mailing by horse, cataloging in a library, using citation statistics, although having large differences in terms of application, technical means, and social roles, by careful analysis, the essence is always the same thing: to meet people's information needs, from low to high, from simple to complex, ranging from production and transmission to storage and processing, from passive to active service provided.

The Internet is also along this path of development. Today, working-links and technical means of the data development have been enriched: digital technology, communications technology, standardization technology, storage technology for enormous quantities, artificial intelligence analysis, software engineering, data sharing technologies, etc.

The mode of data development has progressed from the manual, mechanical, and electronic to network**,** while the explosive growth of information has brought new problems of data integration. The hypertext of the first-generation Internet (web1) WWW only links web sites loosely, forming a fluid, large global data commonwealth, similar to the title catalog of a library, only finding outer clues to data but not accessing the deep ordered substance. Represented by Yahoo and Google, the portal engine and the search engine came in time, just as the classificatory catalog of the library and keyword index to literature, increasing orderliness and

depth of information retrieval significantly. A special example is the Google search engine, with its smart Page-rank technology, according to the links between web pages, generated by classification results in the use of the web-site producers' strength, rather than in the traditional way. The orderly processing of data collections, thus avoids the traditional heavy, backward production and processing methods well adapted to the changing requirements. This new development mode of data is a glimmer of hope.

Traditional data development modes always had such a mark: the post-produced effect lacked real-time with respect to nature. The collection & processing of data for retrieval demands, just as the relationship between theory and life, always appeared as a grey color to lag behind. Because of the limitations of processing speed and storage space, the traditional development modes had to use grey, static sets of standards to regulate and share metadata and data and use photography, reproduction, microforms and other technologies to preserve large-quantity data. Therefore, passive data sharing and the short survival cycle of data cannot be avoided. When a new theory, concept, vocabulary, value, expression**,** image, audio and storage medium arose, even if standards were updated rapidly, a large number of information isolated islands or data wreckages were still left (However, they also required hard work). At least this used to be the case before the network mode came into being.

The Internet era has brought high-speed operations, broadband communications, and a flood of storage. It is not an elusive thing for rapid access to information with its enormous storage and processing. The practice of Google and other search engines has proven this point. On the basis of software engineering and the above technologies, its essence is that for the first time to provide services for obtaining and processing new data whose main features are natural language and non-standard expressions, that is its nature of quasi-real-time effects. This will continue to approach real and immediate effect and provide a complete chain and circuit of data for the steps of collection, storage, processing, transmission and providing, which were separated in the pre-Internet work.

At the same time, this path has laid the foundation for a self-learning mechanism for the integration of data. From the electric circuit to the accumulating thesaurus using type-writers, self-learning mechanisms are already very common in technical applications. It is a novelty, however, at the system level of data development, as before the Internet era, the limitations of computation speed, storage space and aggregation means could only run in a limited local range.

2 THE CHANGING INTERNET

Now, this era is facing four types of data: dynamic data, static data, metadata and standards, and norms. For dynamic data, Google's practices have given the example, but its advanced nature is not consciously considered by enterprises, institutions, and government departments. If a professional system sets up its own data search engines (or other collection mechanism) to collect relevant information, it will play a great role to form the basis for promoting the professional data integration. Particularly in the area of electronic commerce and electronic government administration, much dynamic information will be gradually retained in the system by self-learning systems. Natural language and non-standard data can be put into applications immediately. On the other hand, data can be selected by comparison to standards**,** through the data links or data mining, and then tested in testing database to form standard data, while accumulating material for selection of metadata & development of standards and norms. As for static data, such as the depository for library and backward retrieval data, although

many are not suitable for self-learning mechanisms, but as noted above, the metadata, standards and norms to describe them can be used in collection, analysis, and certification. The relevant databases completed can be also refreshed by self-learning mechanisms. Therefore, we can say, search engines plus self-learning mechanisms should be advanced mode of data development in network era, though its data processing technology is not advanced.

However, the network mode itself is also changing. The underlying motivation is in development and progress itself, which can be seen from two directions from the development trajectory of the Internet:
1) The changes from the evolution in user market demand: if WWW (commonly known as Web1) is the first generation, and the Google search engine can be counted as the 1.5 generation. The current full swing, blog, Wiki, IM, RSS, SNS, is widely known as Web2.0, apparently owing to changes in the users market demand. Being developed and familiar for more than a decade, the first generation of the Internet can no longer meet market demand; users are desperately in need of a direct, proactive, real-time exchange replacing the old indirect, passive, slow manner. This is well known P2P (person-to-person). Because it is the spontaneous emergence of the wave of users, it can be characterized as bottom-up. Viewing these new functions and nature, the most obvious features are in the following aspects:

- The host of data processing: the hosts of data-cell production and processing turn from web sites to a users-information services approach: The approach of data-providing turns from batch-provision of one-to-multi to multi-to-multi and even to direct users randomly, that is P2P. At the same time it turns from passive seeking to active clustering providing;
- Information services functions: data from a simple package search, to download, upload, publishing, storage in clusters, and extends its role further, such as the well-known long-tail attributes;
- Network efficiency: The above changes of Web2.0 result in a client-oriented web and undoubtedly will improve the channels and efficiency of data transmission greatly; and
- Network connectivity: The P2P network operating system SNS is no longer limited to loose web connectivity. We have new Victoria besides Web. Whether from the perspective of technical achievement or social organizations, its emergence with APIs (application program interfaces) is of the utmost importance. Each social unit only needs to carry out simple SNS configurations to establish a workflow P2P network without a server.

The characteristic of *user changes in market demand* is the deeper level of data collection, web data-mining evolving to P2P data-mining. It will raise greater and newer challenges in aspects of issues such as meeting more demands, difficulties of technical handling, intellectual property, etc.

2) The change from evolution of technology: Similarly, E-science and the Grid are being carefully operated popularly in the world scientific community. These will eventually move towards the ideal network (Semantic Web) raised by Tim Berners-Lee who has devised WWW. Its purpose is to rely on the next generation Internet GGG where people can dynamically and transparently share in the distribution of resources in different parts of the Internet, such as large-scale computers, databases, applications, services, etc. It is through standardized middle-ware platforms that organize and mobiliz**e** resources, control data flow, and balance load tasks, providing oriented-services to users. Therefore its characteristic lies in sharing resources and efficiency, which are also the conditions and directions for the mode of data development to consider. Because it is in accordance with the traditional approach to plan, organize and implement, its direction is top-down. Although the current level of E-Science or grid may have a long way to go from Tim Berners-Lee's ideal, the data development is a

prerequisite should be considered. The data warehouse and data-mining technology is the core of its priorities.

Viewed from the perspective of web data mining, two major trends implement data development from two angles as information retrieval (IR) and databases (DB). Web2.0 focuses on IR from the perspective of the quality of user information, and to help users to filter information. And the grid focuses on the integration and modeling of data to support complex data requirements from the DB perspective. Whichever trend is most important, the two goals are the same ---- meet growing user requirements. From their own development requirements, squeezing from opposite directions objectively, these trends have promoted existing modes of data development by difference and mutual-supplementary technical characteristics. It should be said that the network approach of data development has not yet been applied to most of the community, and has raised more requirements and possibilities than the initial network mode:

a) Data collection has a stronger characteristic of real-time and entering a deeper level: that is the characteristic of P2P, such as blog (commonly known as Boke in Chinese) activity, is rife currently at the level of individuals, projects, and units. A large number of relevant data are delivered with timely feedback in an open environment without security. Users can be directly involved in the editing the Wiki Encyclopedia for tan understanding of the latest vocabulary and concepts; Podcasting, the release/subscribe of personal video / audio, also enriches data sources with video and audio; GIS projects can even establish an open architecture to acquire urban geography and commercial geographic data information. The emergence and popularity of SNS enables data flows at the enterprise and industry levels to be more colorful (here we can see the merging of two trends).
b) The transmission speed and frequency of data exchange has become faster: RSS are famous for providing specific subject data regularly; the characteristics of P2P and SNS have also increased the speed of communication.
c) Owning to the volume of data accumulation that grows in a geometric manner, it is more difficult to do data selection and analysis: P2P makes the information explosion of the Web1 period seem simple. The importance of analysis technology has become more pressing, the lack of which is regarded as a shortcoming of the original network mode. Such links and the process of distinguishing, abandoning, selecting, classifying, and indexing of experimental values, images and audio, writings, etc. will be used by data mining technology.
d) The complexity of data storage and retrieval is increasing: The arbitrariness and randomness of P2P divides the resulting data flood into two workflows: non-standard and standard, the former for real-time operations and the latter for standard operations. Thus a large number of heterogeneous databases are produced to adapt to different retrieval needs, which raises new requirements for data warehouse technologies.
e) The level of data sharing is enhanced: the emergencies of the grid and SNS (P2P network operating system) have accelerated the process of data sharing. Also put forward is the new old problem: how to cross the data barriers created by local interests, while taking into account sectional working results, intellectual property rights and individual privacy?

How to make new strategies of data development that combine the characteristics of the two major trends of top-down and bottom-up? They should be based on characteristics of developing information technology. Currently the key to all data linkage is data mining (DM) technology, which has resulted in artificial intelligence technology, language analysis, and expert systems technology. DM plays an important role in data collection, processing, retrieval, and data warehouse technology. Being the core of knowledge discovery (KDD), it involves artificial intelligence, expert systems, machine learning, pattern recognition, statistics, intelligent databases,

knowledge acquisition, data visualization, and other fields. Its task is to create models from data, and their development will determine the mode of data development to a large extent.

Thus, in the background of the network mode of data development, data collection at the P2P level plus data mining and data warehouse technology will become a mode of the data development in the vision of a new generation network (or post-web mode). Because data mining has the functions of data cleansing, intelligence data analysis, incremental maintenance in data warehousing, forecast analysis based on models, relative-rules excavation in serial data, etc, the new mode of data development based on the database will no longer be simply a form of post-produced, but a form in real-time. Data flow from P2P terminals will enter a data pool. After that, the functions of data mining will be implemented, and then the data flow will be entered into a multi-store database to be classified. Then databases, knowledge files, data platforms, and data markets from the ongoing processing data workflow will be generated. Of course, the techniques and strategies of data mining are different when the data flow is generated from different areas, basically as divided into two categories: the Internet and enterprise levels.

The available technical strategy modes of data mining at the Internet level are approximately:
1) Structured patterns: These continue to deepen the original web data mining techniques, which are based on the original techniques (IR, Information Retrieval and IE, Information Extraction), to the P2P level:

- Content mining: from web content mining within unstructured text to text-data mining based on knowledge discovery (KDT); from the technique of Bag of Words to latent semantic indexing; exploring classification mechanisms for text, the relationship between files and models and rules of semi-structured text files, establishing data warehouses, knowledge files, and virtual databases; from data mining methods against data in flat files to conducting a study on multi-store database mining algorithms, etc. The techniques will be updated along with deeper target levels.
- Structure mining: Web structure mining targets super-linked relationships, such as contain, quote, and ultra-subordinate, among web files. The target of structural mining on P2P/SNS is the work-flow data units (files, records, fields, etc.), requiring better algorithms and higher speed.
- Usage mining: current web usage mining reveals new trends in the ways of using record mining, personality mining (modeling for individual user records), and related expansion algorithms. It will become the mainstream in the development of data mining technology.

2) Non-structural patterns: These never develop voluminous and exclusive standardized databases but instead build only a small number of core databases, while developing and relying on data warehouse technology with collaboration of data mining technology. However, its systems demand is very high in the terms of resource sharing and collaboration. This is only suitable for professional systems in which grid technology is quite mature.
3) Regular structural mode: This implements non-real-time batch structural operations in a controllable range; and
4) Partial structural pattern: This only implements structural operation against the core work data flow.

Items 2), 3), 4) above could also form a portfolio that suits different application circumstances.

Enterprise level can be a business, industry, or professional system scope. The strategy patterns of data mining technology differ in such factors as: operational characteristics, work patterns, user needs, data environment,

and system goals. However, at the same time each unit has its own internal and external data sources, not only in competitive market trades systems, such as banking, securities, retail, and telecommunications but also in government, research, management, and other departments. Therefore, the enterprise level must design technical support programs in accordance with their own needs. There are several approaches in general: purchase specific applications models or purchase entire applications programs; hire experts or advisory bodies; train professional technicians to form an internal team.

3 CONCLUSIONS

Whatever kind of program, it is inevitable for professional technicians and data analysts to work together. More importantly, it is necessary to recognize trends by the awareness of decision-makers and to make preparations actively at a strategic level for changes in data development models. While we develop search engines and data mining techniques, resource sharing and data collaboration need to be regarded as a matter of urgency to resolve because future modes of data development in the Internet era cannot be implemented without sharing and collaboration. From word—field—record-- database, to sharing platforms, if the cell of sharing and collaboration is added into each stage, the quality and efficiency of data development will be greatly promoted.

Sharing and collaboration not only can break management barriers but also can avoid the low efficiency of the isolated technical divisions and problems. From the perspective of data development in the Internet era, sharing and collaboration is the third factor to be considered beside market demand and technical updating.

As mentioned above, while the next generation Internet remains a vision, which could be Web2.0 or the grid, or either SNS or DM. Whether at the Internet level or at the enterprise level, it is an inevitable trend for the first generation Internet to change to the next generation. This in turn will inevitably lead to changes in the mode of data development. The two elements - market demand and technological updating - have provided strong supports to the change, and sharing and collaboration should be the third potential factor.. It is important to identify the trends and make preparations at the level of strategy and policy. Thus the pulse of data development in the Internet time can be caught.

4 REFERENCES

Han, J. & Kamber, M. (2001) *Data Mining: Concept and Techniques*. Morgan Kaufmann Publishers, Inc.

O'Reilly, T. *What Is Web 2.0?* Retrieved from the WWW, December 18,2007: http://tim.oreilly.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html

陈文伟, 黄金才编著, 数据仓库与数据挖掘, 人民邮电出版社 (2004)

杨庆跃， Web 数据挖掘的研究现状及发展. Retrieved from the WWW, October 17, 2007: http://www.sunleap.com/personal/article1.htm.