TELKOMNIKA Indonesian Journal of Electrical Engineering
Vol.12, No.6, June 2014, pp. 4631 ~ 4638
DOI: 10.11591/telkomnika.v12i6.5445

An Efficient System for Information Recommendation

Zhenhua Huang*, Qiang Fang
Department of Computer Science, Tongji University
1239 Siping Road, Shanghai, P.R. China
*Corresponding author, e-mail: shtj08.zhh@gmail.com

Abstract

A recommendation system is the one of the most effective tools for tackling with the problem of information overload. However, as the maturity of Web 2.0 and the emergence of massive information, the existing information recommendation systems have the serious drawbacks in the aspects of real-timing, robustness and self-adaptability. Motivated by the above facts, in this paper, we design SIRSCA, which is an efficient semantic-driven information recommendation system under the cloud architecture. Specially, the SIRSCA system mainly include four modules: semantics representation of foundation data and user preference informations; indexing mechanism of massive semantic informations under cloud architecture; recommendation approaches based on semantic computation theory; and technologies of dynamic migration under cloud architecture. We present the extensive experiments that demonstrate our improved system is both efficient and effective.

Keywords*: cloud computing, recommendation system, semantics, dynamic migration*

1. Introduction

The number of servers and webpages accessing to Internet increase exponentially as the in-depth application of information and network technologies. And we need to face the massive information caused by the rapid development of the Internet technology. For example, there are millions of books on Dangdang, millions of films on Netflix, millions of new arrival items on eBay, over fifteen hundred millions of pages on the social network del.icio.us [1]. Information overload has appeared and users can't accurately find their interested items. Also information overload will reduce the economic benefit and market competitiveness for enterprises. According to [2], we know that information recommendation systems is the one of the most effective tools to solve the problem of information overload. These kinds of systems not only are a commercial marketing tool, but also can efficiently improve users' adhesion. According to the report of McKinsey, information recommendation systems provide 47% and 35% products sales for eBay and Amazon respectively [3].

Recently, researchers have focus on the discussion and design of information recommendation methods, because information recommendation methods are the core of information recommendation systems [4]. To our best knowledge, there are three main information recommendation methods: content-based methods [5-7], collaborative filtering methods [8-10], and hybrid methods [11-13]. Table 1 shows the three traditional recommendation methods used in typical information recommendation systems. These existing methods mainly focus on how to build model of user interest to improve the precision of recommendation results. However, as the maturity of Web 2.0 and the emergence of massive information, there are at least three drawbacks about the existing information recommendation systems: (1) since the information of users and products changes dynamically, they need build models repeatedly; (2) due to the opening characteristic of Web 2.0 networks, they are often attacked by some malicious users, and their software modules are often abnormal; (3) they build models according to current user preferences, and don't consider the evolution of user preferences, which will largely affect the quality of information recommendation and the effect of adaptive personalized recommendation.

To solve the above drawbacks of existing information recommendation systems, in this paper, we proposes SIRSCA (Semantic-driven Information Recommendation System under Cloud Architecture), an efficient information recommendation system based on the underlying

Received December 29, 2013; Revised February 28, 2014; Accepted March 17, 2014

knowledges of data and user preference information, and introduces the semantic computation in information recommendation systems. Meanwhile, to improve the real-time and robustness of information recommendation systems, we present an efficient system architecture which is based on the cloud computing platform, and focus on the index mechanism of massive semantic data and the technique for distributed migration. Moreover, we present the extensive experiments that demonstrate our improved system is both efficient and effective.

Table 1. Three Traditional Methods used in Typical Information Recommendation Systems

Methods	Information recommendation systems
content-based recommendation	Personal Web Watcher, AdROSA, SIFT, LyricTime, News Weeder, Google Alerts, etc.
collaborative filtering recommendation	Amazon, eBay, CDNOW, GroupLens, Ringo, Video Recommendation, MovieLens, ACF, SERF, Connotea, Dangdang, etc.
hybrid recommendation	Fab, Daily Learner, CWAdvisor, Quickstep, Foxtrot, OARs, PBTango, Yoda, Open Bookmark, etc.

2. System Framework Overview

As the development of Web 2.0 technology and the emergence of massive information, the existing information recommendation systems have serious drawbacks in the aspects of real-timing, robustness and self-adaptability. To solve these main drawbacks, we design and develop SIRSCA, an efficient semantic-driven information recommendation system under the cloud architecture. The system framework of SIRSCA is shown in Figure 1.

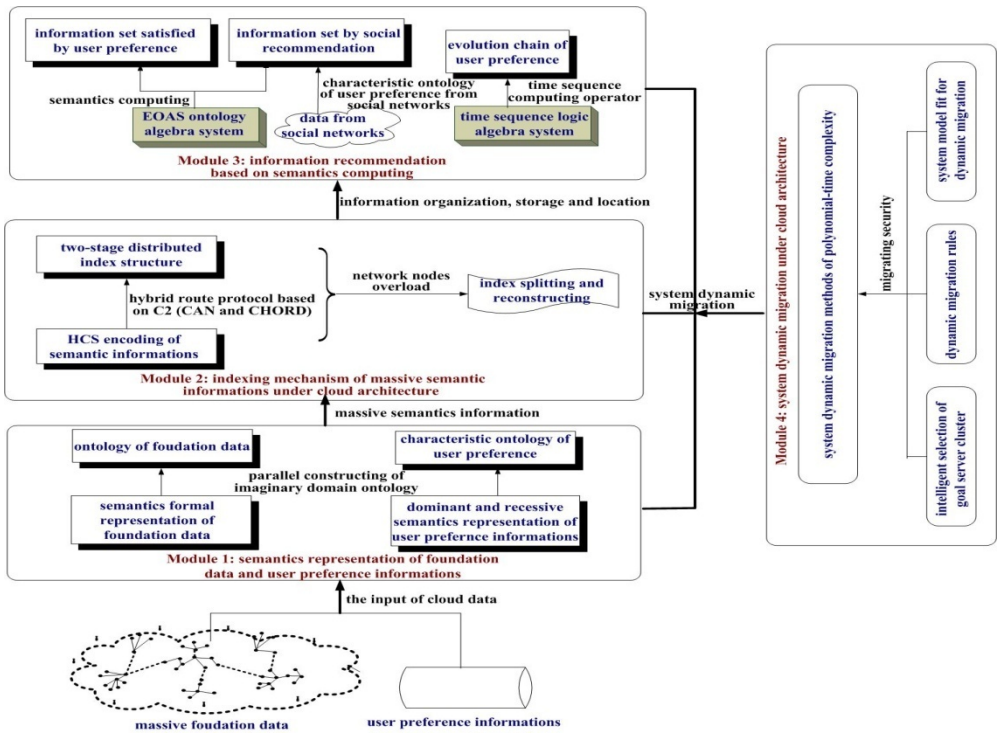

Figure 1. The System Framework of SIRSCA

Our system mainly includes four modules:

Module 1: semantics representation of foundation data and user preference information. In this module, we first define and describe the semantics formal representation of foundation data, and then propose the ontology representation method of foundation data by parallel

constructing of imaginary domain ontology. Meanwhile, we present the dominant and recessive semantics representation of user prefernce informations, and then construct the characteristic ontology of user preferences. Furthermore, we analyse the time and space complexities of the semantic algorithms.

Module 2: indexing mechanism of massive semantic informations under cloud architecture. In this module, we first propose the HCS (Hierarchy Combined Surrogate) encoding of semantic informations for information recommendation systems, and then design a two-stage distributed index structure based on the C2 (CAN and CHORD) hybrid route protocol [9]. Moreover, in order to tackle with the problem of network nodes overload, we present the strategy of index splitting and reconstructing.

Module 3: information recommendation based on semantics computing. In this module, we first design the EOAS ontology algebra system, and then propose the information recommendation methods based on the semantics computing between the ontology of foudation data and the characteristic ontology of user preference. Furthermore, we present the social recommendation mechanism based on the characteristic ontology of user preference from social networks. Finally, we propose the efficient approach for mining the evolution chains of user preference based on time sequence computing operators.

Module 4: system dynamic migration under cloud architecture. Iin this module, we first design the system model fit for dynamic migration and dynamic migration rules. Then we propose an efficient approach for intelligent selection of goal server clusters. Finally, we present the system dynamic migration methods in the polynomial-time complexity with the guarantee of migration security.

3. Specific Realization of Our SIRSCA System

In this section, we give the specific realization of our SIRSCA system.

3.1. Realization for Module 1

Endowing foundation data with understood semantics is the starting point for semantic-driven information recommendation technologies. We find that semantics of foundation data can be defined by the concepts, concept-to-concept relations, attributes, instances and rules in the specific area. In our SIRSCA system, the ontology of foundation data ia defined as O=(C, R, P, I, A), “C” denotes the set of concepts and terms of foundation data; “R” is multivariate mapping from C×C to A; namely, C is the relationship set of concepts; “P” is attribute set of concept features; “I” is the instance set of concepts; “A” is rule sets.

In order to improve the intelligence and the knowledge reuse rate, our SIRSCA system parallel constructs the ontology of foundation data using imaginary domain ontology technology. This method uses the domain description documents of DODL language and the evoluation theory of population of living things (including selection, clone, variation, cross, composing and transgenosis etc.) to combine or delete them. By this way, it can hierarchically process the most fundamental foundation data which initially store in the system, and dynamically construct the global ontology.

As for constructing the characteristic ontology of user preferences, our SIRSCA system uses the idea of the NER (Named Entity Recognition [14]) process in the area of biological knowledge achieving, and employs two-stage modules to express the semantics of user preference information. In the first stage of semantics expression, our SIRSCA system regards the specific user preference informations as a document fragment and use the Latent Semantic Index (LSI) [15] and Support Vector Machine (SVM) [16] technologis to effectively choose concepts of the document fragment. This completes dominant semantics extraction. In the second stage of semantics expression, our SIRSCA system exploits the existing foundation data ontology to find related concepts, relations, attributes, and instances. This completes the latent semantics extraction. Based on the dominant and latent semantics, our SIRSCA system automatically constructs the characteristic ontology of user preferences.

3.2. Realization for Module 2

In the cloud computing environment, the efficiency of information recommendation depends to a great extent on the organization and access mode of massive semantics information. And the distributed indexing mechanism is the one of the most effective approach

for tackling with this problem. Hence in our SIRSCA system, we design the two-level distributed indexing structure C2-DISINX which is based on the C2 (CAN and CHORD) hybrid routing protocol, and uses this routing protocol to appoint specific server clusters for storing corresponding local index. In this way, we can guarantee the scalability and efficiency of the distributed index.

Since the massive semantics information updates frequently, we need to solve the maintenance problem of distributed index. According to the theory of distributed database, the main work of index maintenance is to process the splitting and merging of C2-DISINX index nodes. In our SIRSCA system, we propose an approximate optimal strategy to select the index nodes which need to be split and merged. Let W and S be the set of global and local index nodes respectively. Then our approximate optimal strategy can be described below. We first construct a weighted directed bipartite graph WDBG: map W and S into the vertex set of WDBG, and map the routing cost between nodes which is obtained by sample evaluation to the edge set of WDBG. Then we are based on the shortest path theory and introduce a virtual vertex to convert WDBG to the steiner weighted path graph SWPG in the constant -time complexity. Finally, we produce the steiner tree [17] from SWPG in the polynomial time complexity, and get the approximate optimal solution of index nodes which need to be split and merged. According to the theory of directed steiner tree, the time complexity and the optimization low bound can be adjusted and balanced by the parameter $\partial \in [0, 1]$.

In addition, we find the semantics information is usually massive, if we input index nodes and their data into memory directly, it will cause huge I/O overhead and memory consumption. To solve this problem, in our SIRSCA system, we encode the semantics information by HCS encoding based on Hierarchy Combined Surrogate. The prominent advantage of HCS encoding is that it can improve the efficiency of information recommendation by using less and uniform number of bits to store more data.

3.3. Realization for Module 3

In our SIRSCA system, we use the ONION (ONtology CompositION [18]) ontology algebra theory in the aspect of ontology algebra semantic computation. Since the three operations are set operations, which are lack of quantitative arithmetic arithmetic ability. It is necessary to improve and extend the ONION ontology algebra system because the semantic similarity computation in our SIRSCA system is relied on set operations, arithmetic operations, logic operations and other ones. So we propose EOAS (Extension Ontology Algebra System), which is defined by $\sum=(O, R, Op')$, where $\sum$ is the algebra within ontology, O is the concept set of ontology, R is the four relations (part-of, kind-of, attribute-of and instance-of), Op' contains the set of intersection, union, of difference, also we add the arithmetic operators such as addition, subtraction, multiplication and logical operator. Furthermore, our SIRSCA system proposes a set of sequential computing operator to realize parallel, order, interrupt, recovery, suspend of semantics in the aspect of sequential semantic computing. At the same time, we define the rules such as the laws of calculus, which can ensure correctness of temporal semantic computing.

The essence of information recommendation methods is to find the items which are similar to the description of user preferences and recommends these items to the user. For this observation, we employ the following ideas. The SIRSCA system uses the foundation data ontology and the user preference ontology as the input of information recommendation methods, and utilizes EOAS extension ontology algebra System for semantic computing of the foundation data ontology and the user preference ontology. Then the SIRSCA system achieves retains items that is much similar to the user preference ontology, and discards items that is less similar to the user preference ontology. Specially, in order to effectively integrate the social recommendation mechanism, the SIRSCA system first obtains the preference ontology which is related to the use from social networks, and semantically computes between this ontology and the user preference ontology. Then, the SIRSCA system gets the final recommendation result by taking the semantic computing between the relevance preference ontology and the foundation data ontology.

In order to efficiently achieve the adaptive personalized recommendation, the SIRSCA system proposes the concepts and technologies of evolution chain of user preference. The evolution chain of user preferences is consist of user preference ontologies in different time nodes, which records and tracks changes in different periods of user preferences, and then it

can accurately predict and adjust items interested for the user by analyzing and mining the knowledge in the chain of user preferences.

3.4. Realization for Module 4

An important index of migration methods of information recommendation systems between server clusters is the system outage probability, the downtime, and the function recovery time of soft modules. Therefore, in our SIRSCA system, we needs to solve two technical difficulties: (1) How to select the target server cluster when the system need to be migrated; (2) How to make effective and safe system migration.

For the first technical difficulty, our SIRSCA system uses the query optimizer of existing distributed database management systems to periodically collect meta data by the query optimizer of the existing distributed database management system, including the C2-DISINX indexing mechanism, the value of Hierarchy Combined Surrogate, joint probability or density function of underlying data and repeatability of user preference information. And on this basis, the project chose the note of the minimum cost as the target server cluster of migration.

For the second technical difficulty, our SIRSCA system completes the live migration of information recommendation systems through three stages from the source server cluster NSRC to the target server cluster NDST. In the first stage, our SIRSCA system creates snapshots for the metadata of information recommendation systems, and migrate the snapshots to NDST, and let information recommendation algorithms run on NDST. At the same time, these recommendation algorithms are still running on the NSRC, thus the recommendation results from NDST lag behind NSRC. Hence in the second stage, our SIRSCA system synchronizes the recommendation results of NDST and NSRC circularly. In the third stage, NSRC stops running of the information recommendation algorithms, and copys the different parts of the recommendation results to NDST. In addition, we prove the correctness and effectiveness of the system migration method theoretically.

4. Experimental Evaluation

This section conducts an empirical study of our SIRSCA system using the benchmark synthetic datasets ITEMS and USERS. ITEMS is the set of items which has 20 characteristic attributes, and USERS is the set of users which has 10 preference attributes. We evaluate the efficiency and the scalability of our SIRSCA system.

In the first group of experiments, we fix the cardinality of USERS to 10^6, and let the cardinality of ITEMS vary in the range [1×10^6, 9×10^6]. Figure 2 shows the experimental results for this group.

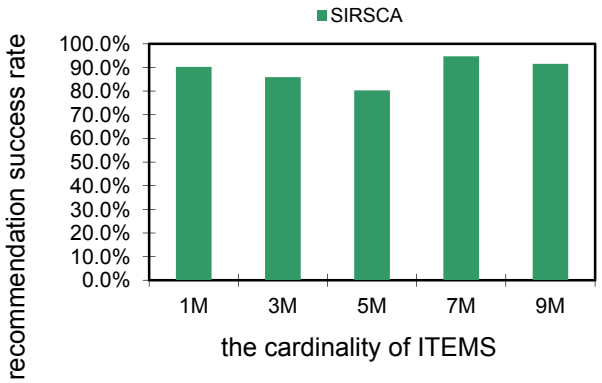

Figure 2. The first group of experiments

In the Figure 2, We can observe that our SIRSCA system has the good implementation performance. Specially, the recommendation success rate of our SIRSCA system in each case is great than 80%. For example, in Figure 2, when the cardinality of ITEMS equals 10^6, the recommendation success rate of our SIRSCA system is equal to 90.2%. And when the cardinality of ITEMS equals 9×10^6, the recommendation success rate of our SIRSCA system is equal to 91.5%.

In the second group of experiments, we fix the cardinality of ITEMS to 5×10^6, and let the cardinality of USERS vary in the range [2×10^5, 1×10^6]. Figure 3 shows the experimental results for this group.

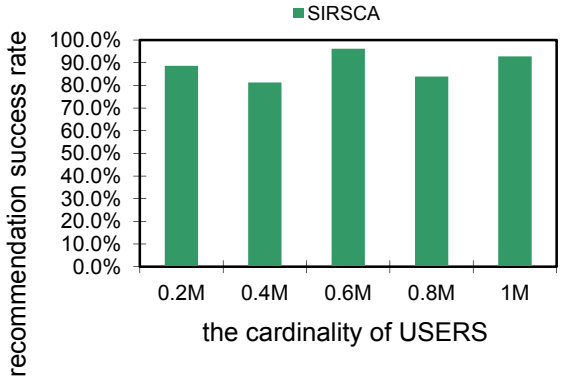

Figure 3. The Second Group of Experiments

In the Figure 3, We can observe that our SIRSCA system has the good implementation performance. Like the first group of experiments, the recommendation success rate of our SIRSCA system in each case is great than 80%. For example, in Figure 3, when the cardinality of USERS equals 2×10^5, the recommendation success rate of our SIRSCA system is equal to 88.6%. And when the cardinality of ITEMS equals 1×10^6, the recommendation success rate of our SIRSCA system is equal to 92.8%.

In the three group of experiments, we let the cardinalities of ITEMS and USERS vary in the ranges [1×10^6, 9×10^6] and [2×10^5, 1×10^6] respectively. Figure 4 shows the experimental results for this group.

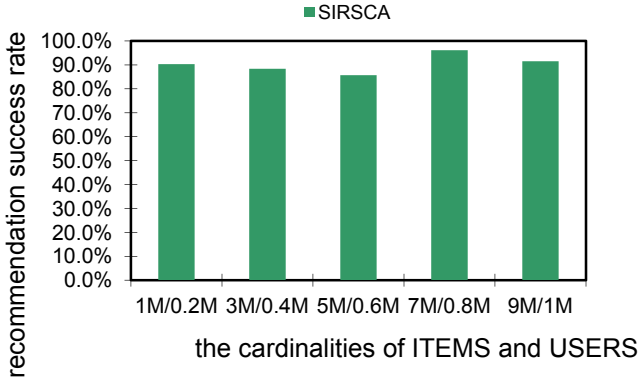

Figure 4. The Third Group of Experiments

In the Figure 4, We can observe that our SIRSCA system has the good implementation performance. Unlike the above two groups of experiments, the recommendation success rate of our SIRSCA system in each case is great than 85% in this group of experiments. For example, in Figure 4, when the cardinalities of ITEMS and USERS equal 1×10^6 and 2×10^5 respectively, the recommendation success rate of our SIRSCA system is equal to 90.3%. And when the cardinalities of ITEMS and USERS equal 9×10^6 and 1×10^6 respectively, the recommendation success rate of our SIRSCA system is equal to 91.5%.

6. Conclusion

Information overload has appeared as the maturity of Web 2.0, and information recommendation systems play important role for mining the potential consumption tendency and finding the items that users are interested in. In this paper, we design the SIRSCA system which is an efficient information recommendation system of new generation and is based on the cloud computing platform architecture, semantic-driven foundation data connotation and user preference. Specially, in our SIRSCA system, we propose four modules: semantics representation of foundation data and user preference informations; indexing mechanism of massive semantic informations under cloud architecture; recommendation approaches based on semantic computation theory; and technologies of dynamic migration under the cloud architecture.Our SIRSCA system drastically changes the status quo that the existing information recommendation systems focus on the mathematical characteristics of data, and ignore the underlying knowledge semantics of data, and provides a novel theoretical and technical way for the information recommendation.

Acknowledgements

This work is supported by the New Century Excellent Talents in University (No. NCET-12-0413), the National Natural Science Foundation of China (No. 61272268), and the Fundamental Research Funds for the Central Universities (Tongji University).

References

[1] Beilin L. *A Study of Personalized Recommendation Evaluation based on Customer Satisfaction in E-commerce*. Proceedings of the International Conference on Computer Science and Service System (CSSS). Nanjing. 2011: 129-132.
[2] Kobayashi I, Saito M. A Study on an Information Recommendation System that Provides Topical Information Related to User's Inquiry for Information Retrieval. *New Generation Computing*. 2007; 26(1): 39-48.
[3] Porat AD. Mass Communication on Social Media: Strategy for Scaling up Personal Conversations. *Journal of Digital & Social Media Marketing*. 2013; 1(1): 74-81.
[4] Liu NH. Comparison of Content-based Music Recommendation using Different Distance Estimation Methods. *Applied Intelligence*. 2013; 38(2): 160-174.
[5] Debnath S, Ganguly N, Mitra P. *Feature Weighting in Content based Recommendation System Using Social Network Analysis*. Proceedings of the 17th International Conference on World Wide Web (WWW). Beijing. 2008: 1041-1042.
[6] Wartena C, Slakhorst W, Wibbels M. *Selecting Keywords for Content based Recommendation*. Proceedings of the 19th ACM international conference on Information and knowledge management (CIKM). Toronto. 2010: 1522-1536.
[7] Cantador I, Bellogín A, Vallet D. *Content-based Recommendation in Social Tagging Systems*. Proceedings of the fourth ACM conference on Recommender systems (RecSys). Barcelona. 2010: 237-240.
[8] Shi J, Long M, Liu Q, Ding G, Wang J. *Twin Bridge Transfer Learning for Sparse Collaborative Filtering*. Proceedings of the 17th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining (PAKDD). Gold Coast. 2013; 7818: 496-507.
[9] Wei S, Ye N, Zhang S, Huang X, Zhu J. *Collaborative Filtering Recommendation Algorithm Based on Item Clustering and Global Similarity*. Proceedings of the 5th International Conference on Business Intelligence and Financial Engineering (BIFE). Lanzhou. 2012: 69-72.
[10] Karatzoglou A, Amatriain X, Baltrunas L, Oliver N. *Multiverse Recommendation: n-Dimensional Tensor Factorization for Context-aware Collaborative Filtering*. Proceedings of the fourth ACM conference on Recommender systems (RecSys). Barcelona. 2010: 79-86.
[11] Burke R. Hybrid Recommender Systems: Survey and Experiments. *User Modeling and User-Adapted*

Interaction. 2002; 2(4): 331-370.
[12] Esfahani MH, Alhan FK. *New Hybrid Recommendation System based On C-Means Clustering Method*. Proceedings of the 5th Conference on Information and Knowledge Technology (IKT). Shiraz. 2013: 145-149.
[13] Chen X, Liu X, Huang Z, Sun H. *RegionKNN: A Scalable Hybrid Collaborative Filtering Algorithm for Personalized Web Service Recommendation*. Proceedings of IEEE International Conference on Web Services (ICWS). Miami. 2010: 9-16.
[14] Nadeau D, Sekine S. A Survey of Named Entity Recognition and Classification. *Lingvisticae Investigationes*. 2007; 30(1): 3-26.
[15] Hofmann T. *Probabilistic Latent Semantic Analysis*. Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence (UAI). Stockholm. 1999: 289-296.
[16] Ha M, Wang C, Chen J. The Support Vector Machine based on Intuitionistic Fuzzy Number and Kernel Function. *Soft Computing*. 2013; 17(4): 635-641.
[17] Byrka J, Grandoni F, Rothvoss T, Sanità L. Steiner Tree Approximation via Iterative Randomized Rounding. *Journal of the ACM*. 2013; 60(1): 1-35.