International Conference on e-Business 2007 (INCEB 2007)
22 - 23 November 2007
P23

A Spreading Activation Approach for e-Commerce Site Selection System

Nilamit Nilas
Faculty of Engineering
Rajamangala U. of Tech.,
Phra Nakhon, Thailand
Email: nilamit@rmutp.ac.th

Phongchai Nilas
Faculty of Engineering
King Mongkut's Inst., of
Tech., Ladkrabang, Thailand
Email: knphongc@kmitl.ac.th

Kasiphan Masakul
Faculty of Informatics
Sripatum University
Bangkok, Thailand
Email: masakul@spu.ac.th

Abstract

This paper presents a dynamic associative network model for e-commerce site selection system based on the psycholinguistic theories of human memory; Spreading Activation Network (SAN). The system is designed to give personalized suggestions based on the user's current personal preferences, other user's common preferences, web-link structure, and activation rules. This work employs a SAN as a technique to provide the evaluation and selection mechanism that provides multiple parallel processes for perception by representing dynamic associations among web-links, user activities, and the relevance subjects of the websites. The system attempts to evaluate a number of sites in an unpredictable complex dynamic environment. Spreading activation explains the predictive top-down effect of knowledge. These processes select the group of the most suitable websites (context) in response to the current conditions (e-commerce activities) while the system continues working towards the user objective goal.

Key Words: Website Monitoring, Customer Behavior, Selection Systems, Collaborative Filtering, and Evaluation e-Commerce.

1. Introduction

An electronic commerce (e-commerce) is a system to facilitate and enhance commercial or transaction through the use of information and communication technology. The e-commerce allows sellers and buyers to communicate, transfer information and interact via computer networks, web-based systems, etc. A typical commercial style is usually based-on market-liked system. However, the commercial has rapidly changed from market-liked commerce to electronic transaction and e-commerce through the online network, Internet, and multimedia device. The e-commerce could cover a spectrum of activities from online product search, to supported sales and services (the Internet, email, discussion forums, and collaborative software). As website and Internet technologies become more established and dependable, the number of Internet users grows fast. In January of 2007, Netcraft, an Internet monitoring company that has tracked Web growth since 1995, reported that there were 106,875,138 Websites with domain names and content on them in 2007, compared to just 18,000 Websites in August 1995.

In recent years, there is a huge increase of electronic commerce and electronic business applications operating over the Internet [1]. Thus, selection systems are increasingly being used in many application settings to suggest products, services, and information items to their customers [2]. For example, many companies such as Amazon.com, Half.com, Dell.com, etc. have successfully deployed selection systems and reported increased Web and Internet sales and improved customer loyalty in their services. Past researches in selection systems have been focused on developing the generic recommendation technologies. Most of the traditional selection systems mainly centers on extracting and recommending the common preferences based on user's historical or preference data [2, 4]. Relatively little work has been done in enabling the system to associate the web-link structures as well as the user current state of activities into the recommendation algorithm and yet be able to recommend the most suitable e-commerce website for him/her. In e-commerce, most recommendation algorithms use the following three types of data: product attributes (product data), consumer attributes (user preferences), and previous interactions between consumers and products, such as buying, rating, and catalog browsing (historical data). Despite the widely

use of this collaborative filtering data, there are several problems limiting its applications. One major problem is that it depends heavily on the product information and the user preference. The system does not have any information of the website or the vender itself. A second problem is lack of understanding of relative strengths and weaknesses of the vender website compare other sites. Moreover, while using traditional selection systems, it is not easy for the users to distinguish whether the items contained in a page are actual recommendations or simply the contents of the page which are displayed indiscriminately to all users. Hence, traditional selection systems do not give the customers the impression of being treated individually. Personalized recommendation agents are emerging to overcome the impersonal nature of integrated recommendations by using technology to assist customers to do decision-makings in treating each customer individually [4, 6].

This paper proposes a framework that employs Spreading Activation Network to analyze and compute the appropriate activation level of each website based on the link structure, the content, the historical data, as well as the preference of the user. This approach enables the system to evaluate a number of sites in an unpredictable complex dynamic environment and recommend the most suitable websites in response to the current context while the system continues working towards the user goal.

2. Website selection system

In this works, we implement the Spreading Activation Network for e-commerce site selection system with four main data types including the common user preference, the product preference, the web-link structure, and the customer/trader historical data. The system combines these data to compose the SAN with the activation number of each competency node to dynamically determine and select the most relevant group of products or websites.

2.1 Common User Preference

The common user preference is a user-centered algorithm that estimates the user's future activities (e.g. transactions, products) by aggregating the observed transactions of similar consumers and common preferences in similar conduct. The algorithm first computes a consumer similarity matrix CS = [csab], s, t =1, 2, …, N. The similarity score csab is calculated based on the row vectors of A using a vector similarity function [4, 5]. A high similarity score csab indicates that consumers a and b may have similar preferences since they have previously purchased common products. The matrix CS·A gives potential scores of the products for each consumer. The element at the x^{th} row and y^{th} column of the resulting matrix aggregates the similarity between consumer and other consumers who have purchased product p previously. The more similar to the consumer is the set of consumers who bought the target product, the more likely the target consumer will also be interested in that product.

2.2 Product Preference

The common product preference is similar to the common user preference. It finds the similarity of the product that could group together. The product similarity level could be calculated using a vector similarity function to form the similarity matrix [5]. This algorithm first computes a product similarity matrix PS = (psst), s, t = 1, 2, …, N. The similarity score psst is calculated based on column vectors of matrix A [1]. A high similarity score psst indicates that products s and t are similar in the sense that they have been co-purchased by many consumers. The matrix A·PS gives the potential scores of the products for each consumer. The element at the c^{th} row and p^{th} column of the resulting matrix aggregates the scores of the similarities between product p and other products previously purchased by consumer c.

The perception behind this approach is similarity of the alternative or the same product group. The more similar to the target product are the products purchased by the target consumer, the more likely customer will also be interested in that product.

2.3 Historical Data

This approach employs probability to determine the patterns of interactions between consumers and products [4]. The interaction matrix A is considered to be generated from the following probabilistic process:

(1) Select a consumer with probability P(c)

(2) Choose a latent class with probability P(z|c)

(3) Generate an interaction between consumer c and product p (i.e., setting acp to be 1) with probability P(p|z).

Thus the probability of observing an interaction between c and p is given by = Σz P(c, p) P(c)P(z | c)P(p | z).

Based on the interaction matrix A as the observed data, the relevant probabilities and conditional probabilities are estimated using a maximum likelihood procedure called Expectation Maximization (EM). Based on the estimated probabilities, P(c, p) gives the potential score of product p for consumer c.

2.4 Web-link Structure

Web-link structure is one of the useful factors that calculate the quality of a page (e.g. page A) proportionally to the quality of the pages that contain inbound-links to it. For example, if the e-commerce site A has an inbound-link from a high quality trader of site B then site A could be considers as a high potential of qualify e-commerce site than other pages that do not have a high quality link from outside.

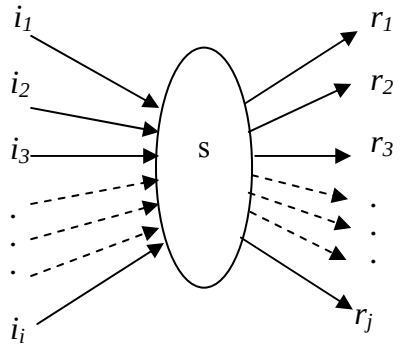

Figure 1. Web-link structure algorithm

The recommendation score of a page is calculated recursively as follows:

$$PageScore(Pn) = \sum_{\substack{at\ page\ i\ to \\ page\ s}} \left(\frac{PageScore(i)}{c(i)} \right) \quad (1)$$

where n is the total number of pages in the collection, and c(i) is the number of inbound link pages i.

As Figure 1 shows, a page p can have a high score if many pages i link to it. The scores will be even higher if the referring pages also have high scores. Our approach is an adaptation of the PageRank [2] algorithm without the damping factor. This method has proved effective in ranking recommendation results. However, the algorithm is computationally expensive because it calculates the score of each page iteratively.

3. Spreading Activation Network

The Spreading Activation Network (SAN) model has its roots in the field of Psychology. This model is the result of studies of the mechanisms of human memory [7]. It has been used in several applications in Computer Science, especially in the area of Artificial Intelligence.

The spreading activation model consists of a network data structure upon which simple processing techniques are applied. The network data structure consists of nodes interconnected by links. Nodes may represent objects or features of objects (websites). The nodes are usually labeled with the names of the objects they represent. The links model the relationships between the objects or the features of the objects. Links may be labeled (web-link) and/or weighted. Links usually have directions, reflecting on the relationship between the connected nodes.

Spreading activation network techniques are iterative in nature. Each network consists of one or more evaluation point (a termination check). The processing is simply a sequence of such iterations, which can be terminated by the user, or by the system. Each evaluation is made up of three stages: pre-classification, spreading and post-classification. The first and third phases are optional. They provide some form of activation decay to be applied to the activated nodes. Thus, the retention of the activation from previous evaluations can be avoided. In this way, there is some form of control over the activation of the nodes in the network.

The spreading phase of the evaluation consists of the flow of activation waves from one node to all other nodes connected to it based on the activation level that compute from the recommendation algorithm. The activation input into a node can be represented by the following simple formula:

$$SA_j = \sum_i \left[(output\ i)(weight\ ij) \right] \quad (2)$$

where SA_j is the total input of node j, (output i) is the output of unit i connected to node j, and (weight ij) is a weight associated to the link connecting node i to node j.

The weight values of the input and weight depends on the application algorithm. In this work, weights have real values indicating the strengths of the association between the nodes. Therefore, there will be four input values for each node based on each method in section 2. After the input value of a node has been computed, the activation of the node is determined. This is given by a function of the input:

$$A_j = f_j(I_j) \quad (3)$$

where A_j is the activation of node j, f_j is the activation function, and I_j is the input of node j. The function f_j takes many different forms based on the method in section 2. This paper computes all the activation score of each method and calculates actual activation value be averaging all activation values from the recommendation scores in section 2.

The output of each node, O_j, is usually its activation value (or activation level). The output value is propagated to all nodes connected to the

active node. Usually, the same output is sent to each node. In this way, the activation spreads evaluation after evaluation. After a pre-classification number of evaluation processes, a termination check is carried out to determine whether the termination condition has been met. If so, the SAN process stops. If not, another series of evaluation continues, followed by another termination check. This cycle goes on until the termination condition is met. The end result of SA process is the activation level of each node in the network at termination time and the group of the top 10 highest activations is the recommendation results.

4. Implementation

There has been an extensive research on applying spreading activation models to the information retrieval and information classification problems [8, 9]. This approach can be distinguished from the other approaches by the fact that it represents queries, terms (keywords), sites, and their relationships as a network of interconnected nodes, thus expanding the matches of a query through new matching terms or items that are matches to the original query itself. The node activation process used in the spreading model starts by placing an activation weight at some starting term (an initial query formulation) or a site retrieved in an earlier search operation. From the initial activation weight (node input) then spreads through the network along the links originating at the starting node. The spreading action first affects those nodes located closest to the starting node and spreads through the network, one link at-a-time. Typically the activation weight of a node is computed as a function of the weighted sum of inputs to that node from directly connected nodes.

The system is implemented using JAVA. The interaction with the internet is handled via HTML and PHP scripts. A very simple graphical user interface was built to handle the user input and the presentation of the retrieved recommendation list to the user.

This approach combines of 4 processes:

1. Download web pages from the internet and store in a database.
2. Classify the downloaded pages of each website into groups of e-commerce cluster (e.g. similar products) by calculating the keyword density of the page and the number of keyword-content phrase ratio. This
3. Compute the recommendation score based on the method in section 2.
4. Use the number from stage 1 to find the actual input value of the website (each node) by averaging the recommendation score to one value.
5. Correlate the input value of each node with the website keyword and content. The website with high input value should have more relevant keyword or content to recommendation domain.
6. Construct the spreading activation network based on the input value and find the activation number of each node.
7. Propagate the activation number from one node to others through the network and recomputed the activation number corresponds to the neighboring node value. The highest activation number will be the recommendation results.

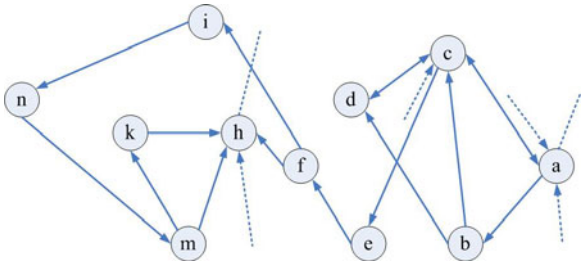

Figure 2. The SAN based on the recommendation scores

In the first stage, the system computes the four input values of each node from each recommendation method in section 2. The node represents the website, thus, the system have four original input values (recommendation scores) for each node. In the second stage, the system calculates the actual node input value by averaging the four original input values from previous stage and continues calculating from one node to other nodes until all nodes has their actual input values.

This stills not the final result of the process yet. In our approach, the system forms a second spreading activation network based on the content and keyword of the e-commerce website. The system uses prior knowledge of the relevant website from database or information retrieval spider.

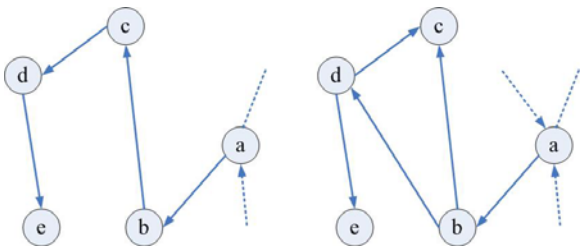

Figure 3. The system correlates node with content and sinks the SAN

This research uses web spiders to apply indexing techniques for text analysis and keyword extraction to help determine whether a page's content is relevant to a target domain. They can incorporate domain knowledge into their analysis to improve the results. The spider checks the keywords on a web page against a list of domain-specific terminology and assigns a higher weight to pages that contain words from the list. Assigning a higher weight to words and phrases in the title or headings is also standard information-retrieval practice that spiders can apply based on appropriate HTML tags.

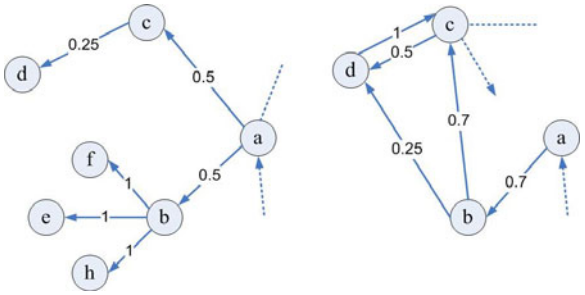

Figure 4. The SAN evaluates the network to select the most suitable nodes

5. Discussion

In this paper, the experiments have been setup to test our approach by first creating a data repository of the relevant e-commerce websites on a local server. Then, the system downloads web pages from the internet and stores them on the server for the evaluation and site recommendation processes. This database serves as the virtual test bed meaning that the system obtains the web page's content, the link structure, and keywords from this local database rather than real-time downloading the web page from the internet during the evaluation process.

The system creates the data repository by start downloading websites from the first five initial sites that inputs by the user. These initials is the starting sites of the relevant e-commerce URL Then, it uses spider to crawl the internet via the web-link of the initial sites randomly. The system downloads the linked sites in a random order. In our experiments, it downloaded 1,000 valid web pages with around 5,000 both internal and external links.

The system also extracts the keywords and the content phrases from each test bed page. It calculated the total number of content phrases of the page and the number of phrases that contained keywords from the keyword list of each page. The web page is considered as a good page if the ratio number was greater than a certain threshold. The ration number is the number of the total content phrases divided by the total number of the phrases that contain website keywords. The system used this keyword-content relationship to classify a set of randomly sampled web pages that will be used to construct the SAN and the recommendation and evaluation process. It found that the keyword-content ratio helps correcting the targeted websites in the SAN and reduce the workload of processing the nod-relevant website. Additionally, the keyword-content ration can be used to attest the recommendation scores of each method in section 2. It also uses to verify the input values of each site (node). If the keyword-content ratio is high and the evaluation score is high, the system would consider the site is more relevant to targeted products than the site with lower ratio or score. The system uses this ratio to classify of the error pages and reduce the total pages in the database.

After classify the downloaded pages, the system computes the recommendation score based on each method in section 2 and calculates the node input score by averaging these scores. Then, the system correlate the node's input value with the website keyword and content. The website with higher input value should have more relevant keyword or content to the recommendation domain. The spreading activation network is constructed based on the input value and find the activation level of each node. The nodes with the highest activation numbers will be the recommendation results.

The execution time of the system is variable depending on the number of nodes on the network. The current prototype implementation of this approach is slow, but efficiency issues were not the foremost consideration for at this stage of the paper. The average time taken for retrieving, indexing and similarity evaluation of one page, the largest part of the execution time is attributed to downloading the website. In the implementation, it requires to verify the recommendation results manually. A team of 5 users were selected randomly. The users were asked to review the recommendation results from the system based on the satisfaction, and usefulness of the recommending results. They reviewed the top 10, 20, 30, 40, 50 and 60 results and provided the satisfaction level (0-1) of these results. The most satisfy result has the higher the number. Figure 5 shows the evaluation from the users.

When reviewing the recommendation results, all of the users satisfy with the website that suggested by the system. However, the satisfaction levels varied depending on the number of the recommending sites. Users provided a very high satisfaction level when reviewed only the top ten results. However, the satisfaction decreases when the number of the reviewing web sites is increase. It could be because of the Spreading Activation Network trends to provide the most suitable node (website) and

recalculates the activation number each time when the node's input value is updated. The system propagates from lower activation node to the higher activation node to find the highest activation number.

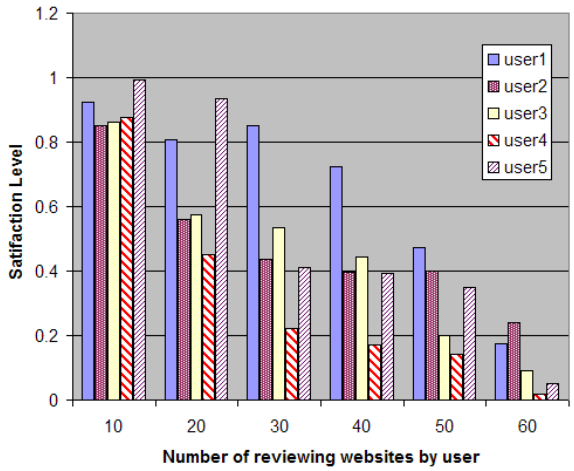

Figure 5. User satisfaction result

Thus, the system propagation and computation would consider the nodes with higher level of activation than the node with middle or lower activation that have the lower keyword-content ratio and lower the recommendation scores.

6. Conclusion

This paper has introduced an approach for e-commerce website selection system with respect to various criteria and features. Additionally, the paper has advocated the need for corroborative system of e-commerce website selection system, eventually presenting SAN as the network model of evaluate the website.

The primary experimental results indicate that our approach is able to recommend strongly positive matches for an e-commerce website based on the product keywords, content of the web, and web-link structure. The preliminary experimental result demonstrates the usability and the capability of the system. This combined capability makes this approach unique and potentially very useful for website evaluation and recommendation. The future work includes performing further experiments with more user evaluation. This evaluation is a need to determine the usability of the designed systems. Future works also include a comparison of this approach with other website selection systems.

7. References

[1] P. Domingos, E. Morgado, "Progressive Rules: A Method for Representing and Using Real-Time Knowledge," Proc., 9th IEEE Conf. on Tools with Artificial Intelligence, 5-8 Nov 1995,pp. 408-415.

[2] M. Chau, H. Chen, "Comparison of Three Vertical Search Spiders," IEEE Computer, May 2003, pp.56-62.

[3] C.N. Ziegler, G. Lausen, "Spreading Activation Models for Trust Propagation," Proc., IEEE Conf., on e-Technology, e-Commerce and e-Service (EEE'04), 28-31 Mar. 2004, pp.83-97.

[4] D. Aswath, et al., "Boosting Item Keyword Search with Spreading Activation," Proc., IEEE on Web Intelligence, WIC/ACM 2005, 15-22 Sept. 2005, pp.704-705.

[5] Z. Huang, "A Comparative Study of Recommendation Algorithms in e-Commerce Application," IEEE Intelligent Systems, 2006.

[6] F. Crestani, P. L. Lee, "WebSCSA: Web Search by Constrained Spreading Activation," Proc., IEEE Forum on Research and Technology Advances in Digital Library, 19-21 May, 1999, pp.163-170.

[7] J. Anderson, "A Spreading Activation theory of Memory," Journal of Verbal Learning and Verbal Behavior 22, pp.261-295, 1983.

[8] T.E. Doszkocs, J. Reggia, and X. Lin, "Connectionist models and information retrieval", Annual Review of Information Science and Technology 25, 1990, pp. 209-260.

[9] F. Crestani, "Application of spreading activation techniques in information retrieval", Artificial Intelligence Review 11, 6(1997), Kluwer Academic Publishers Norwell, MA, USA, pp. 453-482.

[10] P. Nilas, and N. Nilas, "A Dynamic Associative e-Learning Model based on a Spreading Activation Network" IEEE Canadian Conference on Electrical and Computer Engineering, May 2006. pp.2472 - 2475

[11] G. Seralton, and C. Buckley, "On the Use of Spreading Activation Methods in Automatic Information Retrieval", Proceedings of the 11th annual international ACM SIGIR conference on Research and development in information retrieval, ACM Press, NY USA, 1988, pp. 147-160.