문서분류 알고리즘을 이용한 웹 링크 그룹 추천 시스템 연구

문일형, 서대희, 조동섭
이화여자대학교 컴퓨터전보통신학과

Web Link Group Recommend System Design using Page classification Algorithm

Yilhyeong Mun, Dae-Hee Seo, Dong-Sub Cho
Dept. of Computer Science Engineering, Ewha Womans Univ.

Abstract - 본 연구에서는 웹 서비스의 종류가 급격히 증가하게 됨에 따라 유사 패턴의 사용자들을 위해 웹 링크 서비스를 일부 추천해주는 시스템에 대해 설계 및 구현하였다. 본 연구를 통해 유사 패턴의 웹 서비스 이용자들의 그룹을 정의 하는데 네이브 베이지안 알고리즘을 적용하고 그에 따른 새로운 사용자에 대한 그룹정의도 함께 한다. 유사 패턴의 그룹의 사용자들에게 적합한 링크들을 추천해준다. 기존의 추천 시스템에서 제공하는 추천 아이템을 제정의 하는 것이 아니라 기존의 웹 서비스 페이지에서 유사 패턴의 그룹에게만 일부의 링크들만 활성화 하여 제공한다. 이는 웹 서비스의 일부 링크 서비스들만을 활성화 하여 추천 해줌으로써 웹 서비스의 모바일 디바이스등에 제공시 웹 페이지의 소스를 경감하여 좀 더 수월하게 서비스 할 수 있다. 또한 사용자들도 추천 받은 링크만을 접근하게 됨에 따라 접근하지 않는 다른 서비스에 대한 링크 소스가 빠진 웹 페이지만 제공받을 수 있다.

1. 서 론

웹 페이지에 많은 서비스들을 제공되고 있는 가운데 사용자들이 원하는 서비스들 역시 다양해졌다. 그러나 다양해진 서비스들 가운데 사용자들에 맞게 웹 서비스를 추천해주는 시스템들이 필요하게 되었다. 기존의 웹 페이지에서 수많은 서비스들을 많은 사용자들에게 제공하고자 많은 링크 서비스들이 제공된다. 그러나 이렇게 많은 링크 서비스들 가운데 사용자들의 사용 패턴은 한정적일 수밖에 없다. 자신의 관심분야에 국한되기 때문에 불필요한 링크 서비스는 제공되는 것이 의미가 없다. 이런 부분을 이용하여 기존의 추천시스템을 기초로 유사 패턴의 사용자들을 먼저 그룹을 정의하고 새로운 사용자들의 그룹을 정의할 수 있도록 한다. 이렇게 만들어진 그룹에서 필요로 는 가장 알맞은 링크들만을 서비스한다. 이때 네이브 베이지안 문서분류 알고리즘을 적용하고 정확성을 높이기 위해 불확실성 샘플링 알고리즘과 동적임계치도 설계한다.

2. 본 론

2.1 관련연구

2.1.1 HITS(Hyperlink Induced Topic Search)알고리즘[1]

웹 검색으로 연관된 페이지를 찾을 때 다른 페이지를 얼마나 많이 나가는 링크가 있는가의 정도를 허브로 표현할 수 있다. 또한 다른 페이지들이 얼마나 많이 들어오는 링크를 가지는가의 정도를 authorities로 표현하여 높은 권위를 가진 페이지를 찾는 것이다. 즉 아주 높은 권위에 있는 페이지들은 진형적으로 그들 자체로서는 유명하지 않고 많은 수의 관련된 페이지들이 보증인으로부터 얻어진다는 직관에 따라 HITS알고리즘은 시작한다. [3]
질의어와 관계있는 페이지들의 부분집합인 subgraph를 만드는 단계와 만들어진 subgraph를 이용하여 hubs와 권위를 계산하는 단계로 이루어진다.

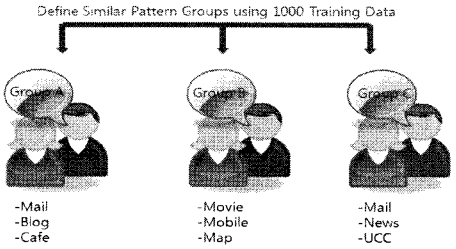
2.1.2 Naive Bayes 알고리즘

확률 이론을 기계학습에 적용한 것으로, 특정 데이터 집합 A를 조사했을 경우, 가서 a가 사실일 확률은 P(a|A)가 된다. 그리고 가설이 사실일 경우 데이터 A의 확률이 P(A|a)일 때, 정리는 P(a|A)=P(A|a)P(a)/P(A)이다. 여기서 P(a)는 데이터에 관한 정보가 주어져 있지 않을 경우 가설이 사실일 사전확률이다. 기계학습에서 중요하게 보는 값은 P(a|A)인데, 베이지안 학습법은 가설집합 H에 포함된 가설 중 최대 확률을 가지는 가설 a를 구하는 것이다.[5]

2.2 링크 그룹 추천 시스템 설계를 위한 사전 유사 사용자 패턴 그룹

2.2.1 학습데이터를 이용한 추천 그룹 정의

기존의 웹 서비스 사이트의 링크 서비스들을 지정된 그룹에 일부만을 제공하고자 한다. 그러기 위해서는 우선 서비스 그룹을 설정한다. 이 서비스 그룹의 설정은 학습데이터를 통해 이루어진다. 1000명의 사용자 학습데이터를 만들고 이를 이용하여 10개 그룹을 정의한다. 그룹의 정의는 인터넷 진흥원에서 제공하는 포탈사용현황 자료를 토대로 임계치를 설정하여 정의한다. 사용현황이 복수 응답으로 이루어진 설문 내용이므로, 사용자들이 70% 사용하는 경우는 링크 서비스가 보다 많은 그룹에 제공되도록 하였다.

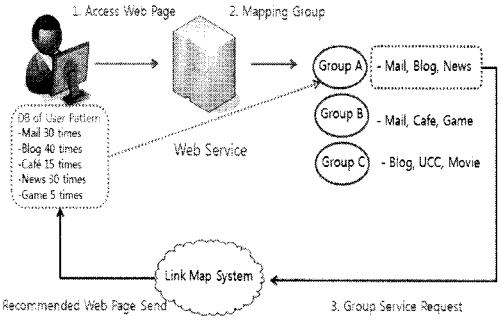


<그림 1> 유사 패턴 추천 그룹 설계

위의 그림에서 알 수 있듯이 각 그룹에 유사한 패턴의 사용자들을 분류하여 그룹을 정의한다. 정의된 그룹에는 패턴에 맞는 서비스들을 분류한다. 각 그룹마다 제공되는 링크 서비스들을 달리하여 제공하도록 한다. 이 그룹 설정은 추후 새로운 사용자들이 링크 추천을 위해 유사 패턴 그룹을 정하는데 필요하다. 그러나 이 추천 그룹 설계에 있어 그룹의 개수와 각 그룹에 제공될 링크의 선정은 가장 중요한 부분이다. 왜냐하면, 사용자들의 요구는 고정되어 있는 것이 아니라 빠르게 변하기 때문이다. 이러한 빠르게 변하는 사용자들의 요구를 최대한 반영 할 수 있도록 추천 그룹의 정의가 이루어져야 한다.

2.2.2 새로운 사용자를 위한 추천 그룹 설정

새로운 사용자가 일정 링크들에 대한 접근이 이루어진 후, 링크 접근 정보들을 이용하여 링크 추천을 위해 추천 그룹을 설정한다. 새로운 사용자는 웹 페이지에 접근하자마자 추천그룹이 설정되는 것이 아니라 일정 링크 접근에 대한 데이터들이 축적된 후, 그 데이터를 이용하여 추천그룹을 설정하게 된다.



<그림 2> 새로운 사용자를 위한 시스템 설계

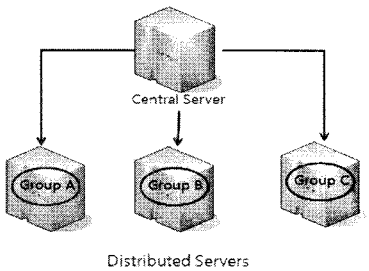
위 그림은 새로운 사용자가 일정 패턴을 알 수 있는 데이터가 생겼을때, 링크 추천을 위한 그룹 설정 및 서비스 흐름을 보여준다. 새로운 사용자가 일정 패턴의 데이터가 데이터 베이스에 생긴 후 웹 서비스를 위해 웹 페이지에 접근한다. 접근한 새로운 사용자를 위해 웹 서버는 이미 정의된 추천 그룹에 새로운 사용자를 설정한다. 데이터 베이스에 있는 사용자 패턴을 분석하여 가장 적합한 그룹에 포함시킨다. 새로운 사용자의 그룹이 정해지면 그 그룹의 추천 링크들이 구현된 웹 페이지를 제공한다. 이 사용자는 추후 추천된 링크가 구현된 웹 페이지를 서비스 받게 된다. 그러나 이런 사용자들의 요구가 장시간 고정되는 경우는 드물다. 즉 많은 사용자들의 요구가 빠르게 변할 수 있다. 이때, 빠르게 변하는 사용자들의 요구에 맞게 유사 패턴의 그룹을 변경해야 하는 문제가 발생한다. 사용자들의 수가 그룹을 변경할 수 있을 정도라면 가능하겠지만 수많은 사용자들의 그룹을 매번 요구를 분석하여 그룹을 재설정하기 어렵다. 그렇기 때문에 최초 그룹 설정이 이루어질 때, 지정된 그룹은 기본 사용자의 속성으로 보고, 사용자들이 다

른 그룹의 링크 추천을 원한다면 다른 그룹의 링크 추천을 받을 수 있도록 한다.

2.3 링크 그룹 추천 시스템을 통한 웹 서버 관리 방안

2.3.1 유사 패턴 그룹을 통한 분산 서버 관리

웹 서비스를 제공하는 공급자는 수많은 사용자들의 관리를 위해 여러 개의 서버를 운영하고 있다. 초기 서비스 웹 페이지를 여러 개의 서버에 구동하게 함으로써, 많은 사용자들을 부하가 적은 서버에 접근시키게 하여 같은 서비스를 제공한다. 이렇게 여러 개의 서버가 관리되는 것을 부하가 적은 서버에 사용자들이 접근하게 하는 것보다 위에서 정해진 유사 패턴의 그룹의 사용자들을 관리하도록 하는 것이 효율적일 수 있다. 즉, 유사 패턴으로 설정된 그룹의 사용자들을 위한 서버를 만들고 그에 따른 링크 추천을 할 수 있도록 한다.



<그림 3> 새로운 사용자를 위한 시스템 설계

위 그림과 같이, 웹 서비스 공급자들은 중앙 서버에 하위 유사 패턴 그룹의 사용자들을 관리하는 분산 서버들을 두어, 웹 서비스의 관리를 효율적으로 할 수 있다. 각 서버는 이미 유사 패턴의 사용자들의 접근을 관리하기 때문에 그 그룹에 서비스 되는 특화된 부분의 링크 서비스들을 관리할 수 있다.

2.4 문서분류 알고리즘을 이용한 사용자 분류

2.4.1 Naive Bayes 알고리즘을 응용한 사용자 분류

Naive Bayes 알고리즘은 위 관련연구에서도 언급되었듯이, 문서분류 알고리즘들 가운데 정확도가 가장 높은 알고리즘이다.[7] 이 알고리즘의 특성을 응용하여 주제를 중심으로 한 하이퍼링크 추천 시스템에 적용하고자 한다.Naive Bayes 알고리즘을 본 사용자분류에 적용하고자 한다. 다음은 사용자분류의 확률을 계산하기 위한 파라미터들과 식이다.

$$P(c \mid x) = \frac{P(c)P(x \mid c)}{P(x)} \qquad (1)$$

$$P(c \mid x) = \sum_{c \in C} \frac{P(c)P(x \mid c)}{P(x)} \qquad (2)$$

위 x는 임의의 사용자를 의미하고 c는 임의의 그룹을 의미한다. 식(1)의 P(x)는 전확률식(Total Probability Formula)에 의해 식(2)와 같이 정의 된다. 식(1)의 분모에 위치한 P(x)와 P(c|x)만 추정하여 사용자 x가 그룹c로 뽑힐 확률을 구할 수 있다. 따라서 이는 모든 사용자들의 수인 |Tu|와 그룹 c에 속하는 모든 사용자들의 수인 |Tu,c|의 비율로 추정할 수 있다. 이는 다음과 같은 식을 성립한다.

$$P(x) = \frac{1}{|T_{u,c}|}, \quad P(s \mid c) = \frac{N_{c,s}+1}{N_c + |T_s|}$$

$$P(x \mid c) \rightarrow p(< S_1, \ldots, S_n > \mid c)$$

Nc는 그룹 c에 있는 모든 서비스들을 말하여 Nc,s는 그룹 c에 있는 서비스들(링크) 중 특정 서비스를 말한다. P(x|c)는 그룹 c안에 있는 서비스들이 독립적으로 존재한다. P(x|c)를 다음과 같은 식으로 표현할 수 있다.

$$P(x \mid c) \prod_{k=1}^{|x|} P(S_k \mid c) \qquad (3)$$

$$P(s \mid c) = \frac{N_{c,s}+1}{N_c + |T_s|} \qquad (4)$$

P(s|c)은 $\frac{N_{c,s}}{N_c}$ 로 간주할 수 있다. 그러나 이 추정치를 식 (3)에 그대로 적용하면, 전체 식의 값을 0으로 만들 확률이 높다. 분류하려는 사용자들 가운데 존재하는 서비스가 확률을 계산하려는 그룹 내에 존재하지 않을 수도 있기 때문이다. 이런 문제를 해결하기 위해 식 (4)와 같이 m-estimate개념을 응용한 기법을 이용한다.

3. 결 론

본 논문에서는 웹 페이지 서비스를 통해 제공 되는 링크들을 이용하여 추천 시스템을 설계하고자 한다. 링크 분석을 통한 연구는 많이 되고 있다. 그러나 이런 링크 분석 기술을 응용하여 유사 패턴의 사용자들에게 요구하는 링크들만 제공한다. 이는 많은 웹 서비스들이 초기 웹 페이지에 수많은 하이퍼링크에 많은 서비스를 제공하고자 한다. 이런 웹 서비스는 웹 페이지의 불필요한 소스를 제거하고 필요한 링크들을 추천하여 사용자들의 웹 페이지 접근을 용이하게 한다. 또한 공급자 입장에서도 유사 패턴그룹을 중심으로 분산 서버를 운용하는 것이 효율적일 수 있다. 단순한 웹 서버의 분산 운영보다는 정의된 그룹을 중심으로 서버가 운용되는 것은 서버마다 속성을 갖고 그 속성에 맞는 관리를 할 수 있다고 생각한다. 따라서 이런 유사 패턴의 그룹의 정의가 가장 중요하다. 그룹을 설정하고 새로운 사용자에게 알맞은 그룹을 지정하기 위해 알고리즘의 다양한 적용 연구가 필요하다. 이런 링크 그룹 추천 시스템 설계에 맞춰 구현을 한 뒤, 시스템에 대한 추천 평가가 이루어져야 할 것이다.

[참 고 문 헌]

[1] Lan Nie, Brian D.Davison, Xiaoguang Qi, 'Topical Link Analysis for Web Search', SIGIR'06, August 6-11, 2006.

[2] Alexander Birukov, Enrico Blanzieri, Paolo Giorgini,'Implicit: An Agent-Based Recommendation System for Web Search', AAMAS2005, July25-29, 2005.

[3] Jon M Kleingerg, ' Authoritative sources in a hyperlinked environment', Journal of the ACM (JACM) archive Volume 46 , Issue 5 (September 1999),pp. 604 - 632 .

[4] Monika R. Henzinger,'Hyperlink Analysis for the Web',IEEE INTERNET COMPUTING, 2001.

[5] Jeffreys J.A., Miller, B.N., Maltz, D., Herlocker, J.L., Gordon, L.R. and Riedel, J.GroupLens,'Applying Collaborative Filtering to Usenet News', CACM,40(3). pp.77-87.1997.

[6] M. Pazzani, D. Bill년, Learning and Revising User Profiles, 'The Identification of Interesting Web sites', Machine Learning 27, Kluwer Academic Puglishers, pp.313-331, 1997.

[7] D.Michie, D.J. Spiegelhalter, and C.C. Tayer, 'Machine Learning, Neural and Statistical Classification', Ellis Horwood, 1994.

[8] Hill, W. and Terveen, L.,'Using Frequency-of-mention in Public Conversations for Social Filtering', CSCW'96, pp.106-112,1996