2000년 한국정보처리학회 추계 학술발표논문집 제7권 제2호

비 감독 학습방법 클러스터링을 이용한 웹 에이전트 효율성 향상에 대한 연구

김지하*, 곽주현*, 김효래**, 이창훈*
*건국대학교 컴퓨터공학과
**경민대학 사무정보자동화과
e-mail : jhsmith@cse.konkuk.ac.kr

A Study on Tools for Agent System Development

Ji-Ha Kim*, Joo-Hyun Kwak*, Hyo-Rae Kim**, Chang-Hun Lee*
*Dept of Computer Science, Konkuk University
**Dept of Automation Of Office Information, Kyung-Min University

요약

클러스터링을 이용해서 문서를 자동으로 분류함으로서 주제별 프로파일을 생성한 후에 사용자의 취향변화에 신속하게 대응할 수 있는 에이전트의 프로파일관리 및 검색관리기법에 대한 연구

1. 서 론

비 감독 학습방식의 웹 에이전트는 사용자의 행동을 분석하여 적절한 문서의 검색을 수행하여야 한다. 이러한 비 감독 학습방식의 에이전트는 사용자에게 직접적인 평가나 입력을 받지 않기 때문에 실제 사용시에 사용자가 검색하는 문서를 기준으로 사용자의 취향을 분석한다. 그러나 사용자가 보는 문서가 여러종류의 다른 주제를 내포하고 있는 경우가 많다. 기존의 프로파일 생성기법을 사용할 경우 이러한 주제들이 하나의 프로파일에 섞여 존재하게 된다. 학술적인 주제와 자신의 개인적인 취미등의 서로 다른 분야들이 하나의 프로파일에 공존한다는 것은 사용자의 취향의 변화에 대해서 적절한 대처를 어렵게 한다.

이러한 문제점은 웹문서의 클러스터링을 이용하여 문서를 분류해 줌으로서 개선이 가능하다. 본 연구에선 웹 문서 클러스터링을 이용하여 여러 주제를 각각의 프로파일로 관리해서 문서를 분류하는 방법에 대해서 살펴보고, 사용자의 취향이나 여러 가지 상황에 따라 변할 수 있는 키워드의 중요도에 에이전트가 어떻게 반응을 하면 사용자가 원하는 최적의 비율로 문서를 가지고 올 수 있는지에 대해서 살펴보고자 한다.

2. 클러스터링[1)]

클러스터링은 산재되어 있는 여러 데이터들에 대한 특징을 찾아서 그 특징에 부합되도록 데이터들을 분류하는 일로 유사한 특성을 가지는 데이터들을 함께 묶어 이들 데이터가 가지고 있는 공통적인 특징을 그 군집의 대표로 나타내어 전체에 산재되어 있는 데이터를 몇 가지의 특성 군으로 나누어주는 것이다.

문서를 분류하기 위해서는 문서들의 특징을 전혀 모르는 상황에서 각각의 문서의 특징을 추출할 수 있어야 하며, 그들 사이의 공통된 패턴을 발견하고, 목

1) 본 연구는 건국대학교 학술 연구비 진흥 사업의 지원하에 수행되었음

적과 일치되는 판별기준을 수식화 할 수 있어야 한다. 본 연구에서 채택한 클러스터링은 다음과 같은 절차를 통해서 수행된다. [1]

1) 단어의 연관성 계산
2) 단어의 병합에 의한 프로파일 생성
3) 문서의 분류

본 연구에선 전체 문서에 출현한 단어의 집합 중에 n번째 단어와 m번째 단어의 연관성을 다음과 같은 방법으로 계산한다.

Rnm = -ln(우연히 두 단어가 문서에서 중복되어 나타날 확률)

이것은 일어날 확률이 높은 사건인 경우

이는 구체적으로

$$R_{nm} = \ln(P_{wc}^{Dc} \times (1-P_{wc})^{D-Dc} \times {}_{D}C_{Dc})$$

Pwc 는 하나의 문서가 두 단어를 동시에 포함할 확률이다. Dc는 두 단어가 동시에 사용된 문서의 수이며 D는 전체 문서의 숫자이다. 이러한 식을 통해서 구해진 단어사이의 연관성을 이용해서 가장 가까운 관계를 가진 두 단어를 중심으로 병합을 수행한다. 병합된 단어의 그룹은 다시 하나의 단어처럼 간주되어 그 다음 병합을 준비한다. 이때 두 단어가 병합된 후에 연관성의 재 계산은 다음과 같다.

$$R_{AB} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} R_{nm}}{n \times m}$$

이때 모든 단어사이의 갱신된 연관성들이 일정 수치 이하로 떨어질 경우 병합을 멈추며 각각의 단어들의 그룹은 프로파일의 후보가 됨으로서 다중 프로파일이 생성된다.

3. 다중 프로파일을 이용한 문서검색비율 관리

본 논문에서의 편의상 A라는 사용자가 있다고 가정하도록 하겠다. 물론 키워드가 여러 가지가 있겠지만 실험을 보다 효율적으로 하기 위해서 각각의 중요도의 성격이 다른 Agent, Network, Data라는 키워드만을 예로 들겠다. 그리고 각각의 키워드에 대한 중요도는 아래와 같다고 생각하여 보도록 하겠다.

키워드 \ 날자	1	2	3	4	5	6	7
Agent	14	24	33	34	43	53	34
Network	33	33	33	33	33	33	33
Data	53	43	34	33	24	14	33
계	100	100	100	100	100	100	100

표3.1 사용자 A의 키워드 중요도

사용자의 입장에서 살펴볼 때 사용자의 환경이나 취향의 변화에 따라서 키워드의 중요도는 달라질 수 있다. 키워드의 중요도는 일정하게 줄어들 수 있고, 일정하게 증가할 수 있고, 일정비율을 유지할 수 있고, 키워드의 중요도가 증가하고 감소하는 것을Random하게 할 수도 있을 것이다.

이러한 사용자의 취향을 반영하기 위하여 A라는 사용자의 중요도에서 Agent라는 키워드는 날자가 지날수록 키워드의 중요도가 꾸준히 증가하고 있고, Network이라는 키워드는 날자가 지남과 상관없이 일정한 비율로 중요도가 지속되고 있고, Data라는 키워드는 시간이 지남에 따라서 키워드의 중요도가 꾸준히 떨어지고 있음을 알 수 있는 키워드를 설정했다.

그렇지만 7일째에는 모든 키워드의 중요도가 비슷한 비율을 나타내고 있다. 이럴 경우 8일째 되는 날 웹 에이전트가 이 사용자의 취향에 맞는 비율로 자료를 가지고 오기 위해서는 어떤 비율로 자료를 가지고 와야 사용자 A에게 맞는 비율로 자료를 가지고 올 수 있게 될 것인가?

이 물음에서 본 논문은 간단한 식 하나를 제안 하고자 한다.

$$[\sum_{i=0}^{n}(\frac{k}{100})x]$$

I : 사용자가 에이전트를 사용한 날자
k : 각 키워드의 중요도
n : 사용자가 에이전트를 사용한 날자 중 최근일
x : 가중치

위의 식을 사용하여 사용자 A의 취향에 맞는 자료를 8일째에 에이전트가 어떻게 가져오는 것이 효율적인지 살펴보도록 하겠다.

먼저 각 날자의 결과에 대한 가중치x가 없는 상태를 살펴보도록 하겠다.

사용자A가 첫째날부터 마지막날인 7일까지 가져온 결과를 보면 전체 키워드 중에서 Agent라는 키워드의 중요도는 14%, 24%, 33%, 34%, 43%, 53%, 33%임을 살펴볼 수 있다. 이 결과를 위의 공식에 대응시켜보면

$$(\frac{14}{100}+\frac{24}{100}+\frac{33}{100}+\frac{34}{100}+\frac{43}{100}+\frac{53}{100}+\frac{34}{100})$$

이라는 결과가 나옴을 알 수 있다. 이 결과를 보면 사용자 A의 Agent에 대한 키워드의 중요도의 총 합은 $\frac{234}{100}$ 임을 알 수 있다.

Network이라는 키워드에 대해서 위의 공식을 적용해보면 다음과 같다.

$$(\frac{33}{100}+\frac{33}{100}+\frac{33}{100}+\frac{33}{100}+\frac{33}{100}+\frac{33}{100}+\frac{33}{100})$$

이 결과의 키워드 중요도의 총합을 구해보면 $\frac{231}{100}$ 임을 알 수 있다.

마지막으로 Data라는 키워드에 대해서 위의 공식을 적용해 보면 다음과 같다.

$$(\frac{53}{100}+\frac{43}{100}+\frac{34}{100}+\frac{33}{100}+\frac{24}{100}+\frac{14}{100}+\frac{53}{100})$$

이 결과의 키워드 중요도의 총합을 구해보면 $\frac{234}{100}$ 임을 알 수 있다.

이를 간단히 설명하기 위해 전체대비 각각의 결과값으로 표현하면 아래의 표와 같다.

	Agent	Network	Data
키워드 중요도	33%	33%	33%

표3.2 가중치 적용전의 키워드 중요도

가중치를 적용하지 않은 결과를 살펴봤을 때 처음에는 관심이 많았으나 시간이 지날수록 키워드의 중요도가 낮아지는 경우와, 처음에는 관심이 적었으나 시간이 지날수록 키워드의 중요도가 증가하는 경우, 시간이 지남과 상관없이 키워드의 중요도가 일정한 경우 모두 중요도의 총합은 같다.

하지만 사용자의 입장에서는 최근에 키워드의 중요도가 크게 나온 키워드의 경우가 더 최근에 많은 관심을 가지고 있는 키워드일 것이다. 이러한 문제점을 해결하기 위해서 각 날자의 결과에 따라서 가중치에 대한 변화를 부여함으로서 최근에 키워드의 중요도가 큰 키워드일수록 키워드의 중요도를 높게 부여하는 방법을 채택했다.

여기에서 주의해야 할 것이 두 가지가 있다.

첫 번째는 사용자가 웹 에이전트를 사용한 마지막날의 가중치를 높게 부여하게 되면 관심이 없었던 키워드라 할지라도 단지 어제 하루 관심이 높았다는 이유만으로 에이전트가 그 전날(1~6일)의 처리결과와 상관없이 그 키워드에 대해서 상당한 중요도를 가진 키워드로 처리할 우려가 있다는 것이다.

두 번째는 반대의 입장에서 사용자가 웹 에이전트를 사용한 마지막날의 가중치를 낮게 부여하게 되면 위의 경우에서 알 수 있듯이 전날의 중요도가 높았다고 하더라도 기존의 중요도가 낮았다고 한다면 그 키워드에 대해서는 상당히 낮은 중요도를 가진 키워드로 처리할 우려가 있다.

그렇기 때문에 일정 가중치를 두어서 키워드에 대한 예전의 중요도와 어제의 중요도를 적절하게 조정해줄 필요성이 있다. 그리고 가중치는 최근의 것일수록 그 가중치의 비중을 높여줄 필요성이 있다. 본 논문에서는 그 가중치 적용 방법에 대해서 날자의 수를 역으로 적용해서 곱해주는 방법을 채택하고자 한다.

위의 사용자 A에 대해서 적용을 해보면 날자가 1, 2, 3, 4, 5, 6, 7일의 순서대로 지남에 따라 역으로 $(\frac{1}{7},\frac{1}{6},\frac{1}{5},\frac{1}{4},\frac{1}{3},\frac{1}{2},\frac{1}{1})$의 순으로 적용을 해주는 방식이다.

이러한 가중치 적용법을 위의 사용자 A의 경우에 적용해 보면

먼저 Agent라는 키워드에 대해서는

$[\frac{14}{100}\times\frac{1}{7}+\frac{24}{100}\times\frac{1}{6}+\frac{33}{100}\times\frac{1}{5}+\frac{34}{100}\times\frac{1}{4}+\frac{43}{100}\times\frac{1}{3}+\frac{53}{100}\times\frac{1}{2}+\frac{34}{100}]$

이라는 결과를 얻을 수 있고 이를 편의상 100분율로 계산하여 더해보면 결과는 아래와 같다.

2% + 4% + 6.6% + 8.5% + 14.3% + 26.5% + 34% = 95.9

같은 방법으로 중요도가 일정비율로 계속해서 적용되어진 경우인 Network이라는 키워드에 대해서도 살펴보도록 하겠다. 이 경우에 대해서는

$(\frac{33}{100}\times\frac{1}{7}+\frac{33}{100}\times\frac{1}{6}+\frac{33}{100}\times\frac{1}{5}+\frac{33}{100}\times\frac{1}{4}+\frac{33}{100}\times\frac{1}{3}+\frac{33}{100}\times\frac{1}{2}+\frac{33}{100})$

이라는 결과를 얻을 수 있고 이를 편의상 100분율로 계산하여 더해보면 결과는 아래와 같다.

4.7% + 5.5% + 6.6% + 8.2% + 11% + 16.5% + 33% = 85.5

마찬가지 방법으로 중요도가 계속 낮아지다가 실험 마지막날에 중요도가 높아져서 다른 키워드와 비슷한 수준에까지 이른 Data라는 키워드에 대해서도 살펴보도록 하겠다.

$[\frac{53}{100}\times\frac{1}{7}+\frac{43}{100}\times\frac{1}{6}+\frac{34}{100}\times\frac{1}{5}+\frac{33}{100}\times\frac{1}{4}+\frac{24}{100}\times\frac{1}{3}+\frac{14}{100}\times\frac{1}{2}+\frac{33}{100}]$

이라는 결과를 얻을 수 있고 이를 편의상 100분율로 계산하여 더해보면 결과는 아래와 같다.

7.5% + 7.1% + 6.8% + 8.2% + 8% + 2% + 33% = 72.6

이를 상대 비교 해보기 위해서 전체 합에 대한 키워드 각각의 중요도를 표로 나타내면 다음과 같다.

	Agent	Network	Data
키워드 중요도	37.8%	33.6%	28.5%

표3.3 가중치 적용후의 키워드 중요도

그럼으로 이 실험이 끝난 다음날에 웹 에이전트는 위의 중요도대로 Agent에 대한 자료는 37.8%, Network에 대한 자료는 33.6%, Data에 대한 자료는 28.5%만큼의 자료를 웹 상에서 가지고 옴으로 인해서 사용자 A에 적합한 비율로 자료를 가지고 올 수 있다.

위의 실험을 통해서 세 가지 경우에 가중치를 부여함으로 인해서 가중치를 부여하지 않았을 때 발생한 다음의 문제점들을 적절히 처리해 줄 수 있음을 볼 수 있다.

① 가중치가 계속해서 올라가다가 실험 마지막날에 가중치가 떨어져서 일정한 비율에까지 떨어진 키워드에 대해서는 상대적으로 가중치를 높여줌으로 인해서 다른 키워드와의 차별성을 줄 수 있다.

② 가중치가 계속해서 떨어지다가 실험 마지막날에 가중치를 높여주어서 일정한 비율에까지 올라간 키워드에 대해서는 상대적으로 가중치를 낮혀줌으로 인해서 다른 키워드와의 차별성을 줄 수 있다.

4. 결론 및 향후과제

본 논문은 클러스터링을 이용하여 자동으로 분류된 웹 문서를 어떻게 효율적으로 관리할 것인가에 대해 다중 취향에 대한 검색 비율관리 방법에 대해 연구하였다. 검색 에이전트가 문서를 어떠한 비율로 가져와야 사용자에게 적합한 비율로 효과적으로 가져올 수 있는지에 대해서 간략한 실험을 통해서 살펴보고 적절하다 생각되는 방법을 제안하였다. 이는 매우 간단하나 일정 기간내에 사용자의 취향의 변화에 매우 빠르게 대처할 수 있다. 그러나 아직까지 장기적으로 가끔 검색되는 문서나 주기적으로 사용자의 관심이 변하는 문서의 경우 이러한 사용자의 주기를 예측하여 미리 제시한다는 것은 매우 복잡한 문제이며 앞으로의 당면과제이다. 이를 위해서는 방대한 사용자들의 검색 패턴에 대한 데이터수집이 선행되어야 하겠다.

5. 참고문헌

[1] 신진섭 "웹 문서 분류를 위한 단어의 연관성 모델과 클러스터링 모델" 2000년 2월 박사학위

[2] 박재균 "웹 에이전트에서 행위기반 학습에 근거한 사용자 관심도 예측에 관한 연구" 2000년 2월 석사학위 논문

[3] John Davies, Richard Weeks, Mike Revett & Andy McGrath, "Using Clustering in a WWW Information Agent". URL : http://www.labs.bt.com/jasper/html/jasclus2.htm

[4] Hartigan,J. "Clustering Algorithms", Wiley, AY, 1975