소셜지수와 질의패턴의 상관관계 분석을 통한 검색 편의성 향상

안무현, 박건우, 이상훈
국방대학교 전산정보학과
e-mail:amh2793@naver.com

Improvement of Retrieval Convenience through the Correlation Analysis between Social Value and Query Pattern

Moo-Hyun Ahn, Gun-Woo Park, Sang-Hoon Lee
Dept of Computer Science & Information, Korea National Defense University

요 약

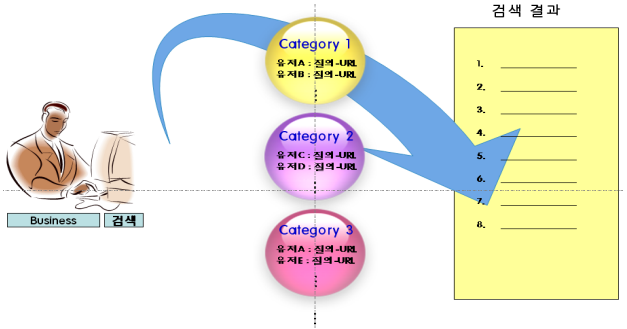
정보의 양이 폭발적으로 증가함에 따라 웹 사용자가 원하는 적합한 데이터를 찾아내는 것은 매우 어렵다. 이는 웹 사용자마다 서로 다른 검색의도와 질의의 모호성에 의한 것으로, 이와 같은 검색의 어려움을 해결하기 위해 많은 연구들이 수행되어 왔다. 질의 로그는 검색자의 검색 의도가 내포되어 있는 중요한 자료이다. 따라서 웹 사용자별 질의 로그 패턴을 분석하여 유사한 질의를 사용하는 웹 사용자들을 클러스터링 하여 검색에 적용한다면 좀 더 유용한 정보를 획득할 수 있다. 즉, 특정 카테고리와 연관된 질의를 자주 사용하는 웹 사용자들은 해당 분야에 관심이 많을 것이며, 또한 다른 카테고리에 관심이 높은 사람보다 상호간에 소셜지수가 높게 나타날 것이다. 특정 주제에 대해 검색을 할 경우 해당 분야에 관심이 높은 웹 사용자들의 질의 및 클릭한 URL 정보를 상속받을 수 있다면 찾고자 하는 정보에 보다 빨리 접근할 수 있다. 따라서 본 연구는 질의패턴 분석을 통해 카테고리별로 관심도가 높은 웹 사용자들을 클러스터링 한 후 해당 카테고리에 대한 정보 검색시 이들이 사용한 질의와 클릭한 URL 정보를 웹 사용자들에게 제공해줌으로써 정보검색의 편의성을 향상시키기 위한 방안을 제안한다.

1. 서론

UC 버클리 대학의 연구 결과에 따르면 새롭게 생성되는 정보의 양은 매년 두 배씩 증가할 것이라고 한다. 우리는 검색엔진이 엄청난 데이터의 홍수 속에서 양질의 정보를 찾는데 도움을 줄 수 있을 것이라고 생각하지만, 검색엔진은 단순히 더 많은 데이터의 존재만을 우리에게 알려줄 뿐이다. 즉, 데이터의 양이 증가하면 할수록 그 속에서 양질의 데이터를 찾아내기는 더욱 더 힘들어 질 것이다.

본 연구에서는 질의 로그 분석 및 검색패턴과 소셜지수의 상관관계 분석을 통해 좀 더 효율적인 정보검색 방안을 강구하고자 한다. 질의 로그는 검색자의 검색 의도가 포함되어 있는 중요한 자료이다[1]. 하지만 질의의 모호성으로 인해 단지 검색엔진 창에 질의를 함으로써 원하는 정보를 찾기란 쉽지 않다. 이때, 특정 분야에 관심을 가지고 있는 집단이 있다면 이들이 사용한 질의와 URL 정보는 관련 분야를 검색하고자 하는 사람들에게 중요한 가치를 가질 것이다. 이는 검색하고자 하는 정보가 이들 집단에서 자주 방문한 URL에 존재할 가능성이 더 높기 때문이다. 즉, 해당 주제에 대한 관심이 높은 만큼 관련분야에 대해 다양하고 핵심이 되는 질의를 사용하여 연관성이 높고 적합한 URL에 접근하기 때문에 관련 정보를 제공받는다면 검색의 만족도는 그만큼 증가할 것이라는 것을 예측할 수 있다.

본 연구에서는 특정 주제와 관련된 용어를 질의에 가장 많이 사용한 웹 사용자를 찾고 이들의 검색패턴을 활용함으로써 검색의 편의성을 향상 시키고자 하였다.



(그림 1) 검색엔진 작동 개념도

즉, (그림 1)에서와 같이 어떠한 주제에 대해 웹 사용자별로 관심사가 유사한 사람들을 클러스터링하여 그들이 가지고 있는 정보(질의, URL)를 활용함으로써 검색의 편의성과 적합성을 향상 시키고자 한다. 또한, 카테고리별 유사한 질의를 사용하는 사람들의 소셜지수 비교분석을 통해 유사 질의를 사용하는 사람들끼리 소셜지수가 어떠한 형태로 나타나는지 분석해 본다. 이는 질의패턴과 소셜지수 간의 상관관계 분석을 통해 소셜 지수를 검색에 적용하기위한 중요한 지표를 산출하기위한 목적에서 수행한다.

2. 관련 연구

2.1 문서 클러스터링

클러스터링이란 주어진 데이터를 의미 있는 그룹으로 분류하는 방법으로 문헌검색, 패턴인식, 경영과학 등에 널리 응용되고 있으며 문서 클러스터링은 대용량의 문서를 주제에 따라 분류하는 것이다. 문서 클러스터링은 검색효과와 능률을 향상시키기 위한 목적을 갖는 문헌 집단을 생성하는데 이용되며 문서의 유사도를 측정하는 방법은 다이스 계수(Dice's Coefficient), 자카드 계수(Jaccard's Coefficient), 코사인 계수(Cosine Coefficient) 등이 있다. 이 중 코사인 계수는 값에 상관없이 문서와 문서 간에 가중치 차가 일정한 경우 같은 유사도를 나타내는 단점이 있어 클러스터링을 하기위해 Dice 계수를 사용한다.

특정 카테고리별로 유사한 문서를 자동으로 클러스터링하는 것과 질의를 분류[2, 8]하는 기존의 연구는 많았으나, 특정 카테고리와 관련성이 높은 질의를 사용한 사람을 클러스터링 하여 검색에 활용하고자 하는 연구는 거의 없었다. 따라서 본 연구에서는 특정 기간 동안 사용된 웹 사용자별 질의 셀을 각각의 문서로 가정하고 이들 간에 클러스터링 작업을 수행하였다.

2.2 Social Network

사회 구성원 간 관계를 맺고 있는 구조나 관계망인 인간관계망은 최근 온라인을 중심으로 하여 Social Network의 개념으로 소개되고 있다. Social Network은 가치, 비전, 아이디어, 친목, 성적인 관계, 친족 관계, 혐오, 논쟁, 거래 등과 같은 하나 이상의 상호의존적인 특정한 형태에 의해 묶여진 개인 또는 조직들 단위의 노드들로 이루어져 있다. Social Network Analysis에서 노드는 네트워크 내의 개별 참가자이고, 참가자 사이의 관계를 타이(tie)라고 한다. 노드들 사이에는 많은 종류의 타이(tie)가 있을 수 있다. 학문적으로 수많은 연구에서, Social Network은 가족으로부터 국가에 이르기까지 단계 까지 많은 레벨에서 작용하며 문제 해결의 방법 결정, 조직 관리 및 개인 목표 성취를 위해 중요한 역할을 한다. Social Network에서 노드들 간의 유사성, 친밀성, 보상성, 접근성 등을 소셜지수(Social Value)라고 하고, 이 값이 높다는 것은 노드 간에 유사성, 친밀성, 보상성, 접근성 등의 사회성이 높다는 의미를 가진다. 본 연구에서는 특정 분야별로 클러스터링 된 웹 사용자들의 소셜지수 값을 계산하는 알고리즘을 제안하고 이 알고리즘을 통해 계산된 결과 값이 클러스터 된 집단에 속하는 웹 사용자와 그렇지 않은 웹 사용자와 어떤 차이를 보이는지 분석한다.

3. 질의패턴에 따른 웹 사용자 클러스터링 및 소셜지수 산출

3.1 질의기반 웹 사용자 클러스터링

3.1.1 질의 로그 전처리

불용어 제거란 검색에서 변별력이 매우 낮은 단어들을 제거하는 것으로 본 논문에서는 "Onix Text Retrieval Toolkit API Reference"에서 사용된 가장 보편화된 571개의 불용어를 사용한다[5]. 형태소 분석을 통해 웹 사용자별 질의 셀의 용어 사용빈도를 추출한 후 배열형태의 매트릭스로 나타낸다. 이때 단순 질의만 존재하고 클릭 정보가 존재하지 않는 질의 정보들은 제거하였고 웹 사용자의 검색 의도에 보다 적합한 질의[1]만을 사용하였다.

3.1.2 카테고리별 관련 용어 셀 작성

특정 카테고리와 관련된 용어를 자주 사용하는 웹 사용자들을 클러스터링 하기위해 특정 카테고리와 연관성이 높은 키워드 셀을 작성한다. 카테고리는 ODP(Open Directory Project)에서 분류한 15개의 범주로 하며 여기서 제공되는 하위 디렉토리를 (그림 2)에서와 같이 관련 용어로 사용하였다. 또한, 세부 카테고리별로 관련 용어를 수집하여 사용한다면 하위 카테고리에 대한 상세 질의패턴 정보(질의, URL)를 얻을 수 있다. 본 연구에서는 우선적으로 Sports 분야에 대해서만 연구를 수행하였다.

순번	주제	관련 용어
1	Arts	sing, theater, movie, designer, animation, theater, animation, dance, ….
2	Business	bank, auction, transaction, trade, …….
3	Computer	internet, chatting, computer, P2P, JDBC, JavaScript…….
4	Games	simulation, star-craft, lineage, …….
5	Health	health, fitness, wellbeing, …….
6	Home	…….
7	Kids & teens	…….
8	News	…….
9	Recreation	…….
10	Reference	…….
11	Regional	…….
12	Science	Algebra, Chemistry, Economics, Geography, Geology, Mathematics…….
13	Shopping	…….
14	Society	…….
15	Sports	Sport, soccer, basketball, baseball, golf, ball, swimming, tennis, …….

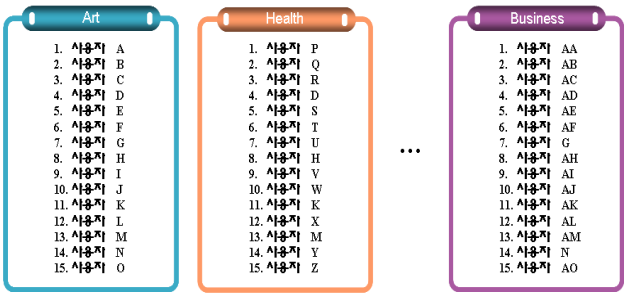
(그림2) 카테고리별 관련 용어 셀

3.1.3 웹 사용자 클러스터링

문헌(카테고리-웹 사용자)간 유사도는 수식 (1)과 같이 두 문헌벡터 사이의 다이스 계수(Dice's Coefficient)로 산출하였다.

$$S(d_x, d_y) = \frac{\sum_{i=1}^{n}(w(t_i, d_x) \times w(t_i, d_y))}{\sum_{i=1}^{n} w(t_i, d_x)^2 + \sum_{i=1}^{n} w(t_i, d_y)^2} \quad (1)$$

카테고리별 웹 사용자 클러스터링 결과는 (그림 3)과 같다.

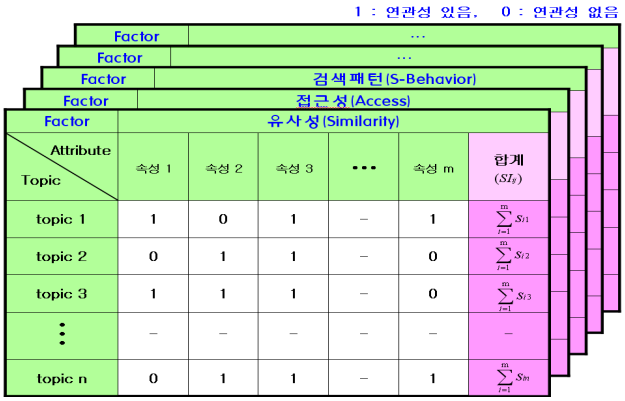


(그림3) 카테고리별 웹 사용자 클러스터링

3.2 소셜지수 산정

3.2.1 소셜지수 산정 알고리즘

소셜지수는 소셜 네트워크에 존재하는 웹 사용자 사이의 잠재적인 특성 및 성향의 관계정도를 타나낸다. 웹 사용자들의 소셜지수가 검색패턴에도 영향을 미치는지 알아보기 위해 아래에 제안한 알고리즘을 이용하였다. 즉, 특정 분야에 대한 관심사가 높은 웹 사용자들의 소셜지수와 서로 다른 분야에 대해 관심사가 높은 웹 사용자들의 지수를 상호 비교함으로써 두 요소사이의 상관관계를 분석한다.



(그림4) 토픽과 소셜지수 속성의 연관성 매트릭스

본 논문에서 제안하는 주제별 웹 사용자간 소셜지수 산출을 위한 알고리즘은 수식 (2)와 같다.

$$SRR_{ij} = \alpha\frac{\sum s_{ij}}{SI_{ij}} + \beta\frac{\sum a_{ij}}{ACC_{ij}} + \ldots\ldots + \gamma\frac{\sum b_{ij}}{SBEH_{ij}} \qquad (2)$$

- SRR_{ij} : 웹 사용자 i, j 간의 Social Relation Ranking Value
- SI_{ij}, ACC_{ij}, $SBEH_{ij}$: $topic$과 관련성을 가지는 유사성($Similarity$), 접근성($Access$), 검색패턴($Search\ Behavior$)의 속성들 합
- s_{ij}, a_{ij}, b_{ij} : 웹 사용자 i, j 간 유사성($Similarity$), 접근성($Access$), 검색패턴($Search\ Behavior$)의 속성 일치 항목
- $\alpha+\beta+\ldots\ldots+\gamma=1(\alpha,\beta,\ldots,\gamma$: 가중치)

토픽에 따라 세부 속성의 관련성 여부는 모두 연관이 있는 것으로 가정하였으며 Factor별 가중치는 인공신경망을 이용하여 최적의 값을 산출하였다.

4. 실험 및 결과

4.1 실험 데이터 구성

사용자 질의는 셀은 AOL(America Online)에서 '06년 3월부터 5월까지 실질적으로 웹 사용자가 입력한 질의를 사용하였다. 공개된 질의어는 약 650,000명으로부터 36,000,000여개 이며, 이 중 24,000명의 질의 셀을 샘플링하여 질의 패턴을 분석하였다.

본 연구에서는 ODP 15개 카테고리 중에서 샘플로 Sports 부분에 국한하여 실험하였으며 관련된 용어는 ODP에서 제공하는 Sports 카테고리 내의 세부 카테고리에 해당하는 항목 398개중 가중치가 높은 140개 항목의 용어들을 사용하였다.

4.2 카테고리별 웹 사용자 클러스터링 및 검색결과

Sports 분야와 관련된 용어 140개를 하나의 문서로 가정하고 웹 사용자별 질의 셀을 각각의 문서로 가정하여 관련 용어가 포함된 질의 정보만을 추출하여 이들 빈도수를 매트릭스로 나타내었다. 이후 Dice 유사도 함수를 이용하여 Sports 카테고리와 가장 관련 있는 질의를 많이 한 상위 50명을 뽑아내어 이들이 Sports 분야에서 가장 많이 사용한 검색어 및 URL 정보를 (그림 5)과 같이 순위화하여 나타내었다. (그림 6)은 "Golf" 용어를 질의에 자주 사용한 상위 50명의 정보를 기반으로 추출된 추천 URL이다.

(그림5) 카테고리 Sports에 대한 추천 용어 및 URL

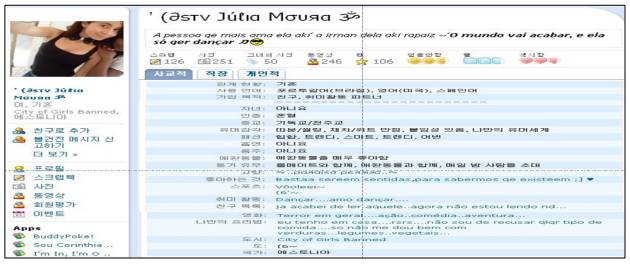
(그림6) 검색어 Golf에 대한 추천 URL

이들 정보는 카테고리 검색엔진을 제공하는 야후에서 추천해주는 사이트와 일부 유사하지만 이 분야에 관심사가 높은 사람들이 제공해주는 정보라는 점에서 의의가 있다.

4.3 질의패턴과 소셜지수 상관관계 분석

소셜정보를 제공해주는 웹 사이트에 방문하여 관심사가 Business, Science, Games 분야인 인물을 각각 10명씩 추출하여 이들의 소셜정보를 활용하여 소셜지수를 비교 분석하였다. 여기서, 각 분야에 대한 샘플링 된 웹 사용자들은 해당 분야별로 가장 활동성이 높은 커뮤니티의 운영자들이다.

전제조건으로, 카테고리별로 클러스터링 된 웹 사용자 중 상위 랭크자는 해당 카테고리에 대한 관심사가 높을 것이라는 가정 하에 (그림 7)과 같은 정보를 제공해주는 orcut.com에서 카테고리별로 활동성이 높은 사람을 선정하여 이들에 대한 소셜정보를 사용하였다.



(그림7) 소셜정보 제공 사이트(orcut.com)

본 논문에서는, 웹 사용자들의 유사성(Similarity), 접근성(Access)을 Factor로 하였으며 유사성에 대한 세부 속성은 성별·결혼여부·나이·언어·지역·학교·직업·관심분야·전공으로 하였고, 접근성에 대한 세부 속성은 동호회·친구등록 여부로 하였다. 토픽에 따라 세부 속성의 관련성 여부는 모두 연관성 있는 것으로 가정(모든 값 = 1)하였으며 Factor별 가중치도 따로 산정하지 않고 동일하게 부여하여 지수를 산출하였다.

소셜지수 분석결과 특정 카테고리로 클러스터링 된 웹 사용자들의 평균 소셜지수가 다른 카테고리로 클러스터링 된 웹 사용자들과의 평균 소셜 지수보다 높게 나타났다. Business, Science, Games에 관심이 높은 사람들의 평균 소셜지수는 각각 0.23, 0.26, 0.21를 나타내며, Business-Science, Business-Games, Science-Games의 평균 소셜지수는 각각 0.18, 0.20, 0.17로 나타나는 것을 확인하였다.

Business	성별	결혼여부	나이	사용언어	사는곳	고교	대학	직종	관심분야	전공	동호회	친구등록
사용자 A	남	미혼	26	영어/한두	인도	Nimala	-	-	비즈니스	-	기준	기준
사용자 B	남	미혼	28	포루투칼	브라질	-	UFMG	-	비즈니스	자동차공학	NO	NO
사용자 C	남	기혼	29	포루투칼	브라질	-	-	하이테크	비즈니스	-	NO	NO
사용자 D	남	미혼	30	영어/한두	인도	BMHS	UTV	소프트웨어	비즈니스	전자공학	NO	YES
...	...	...	...	...	...	...	...	...	...	...	...	...
사용자 E	남	미혼	26	아랍어	이란	Danesh	KNT	하이테크	비즈니스	산업공학	NO	NO
											평균지수	0.23
Science	성별	결혼여부	나이	사용언어	사는곳	고교	대학	직종	관심분야	전공	동호회	친구등록
사용자 AA	남	미혼	32	포루투칼어	브라질	-	-	-	과학	-	기준	기준
사용자 AB	남	미혼	28	영어/한두	인도	-	-	-	과학	-	-	NO
사용자 AC	남	미혼	20	영어/한두	인도	MBI	-	교육	과학	-	-	NO
사용자 AD	여	미혼	24	영어/한두	인도	UPS	-	-	과학	-	-	NO
...	...	...	...	...	...	...	...	...	...	...	...	...
사용자 AE	남	기혼	38	영어	미국	GJ	-	교육	과학	약학	-	NO
											평균지수	0.26
Games	성별	결혼여부	나이	사용언어	사는곳	고교	대학	직종	관심분야	전공	동호회	친구등록
사용자 BA	남	미혼	25	영어/한두	인도	BVB	-	하이테크	게임	-	기준	기준
사용자 BB	남	미혼	31	영어/한두	인도	-	-	학생	게임	응용전자	YES	NO
사용자 BC	남	미혼	32	영어	미국	MBI	-	금융	게임	-	NO	NO
사용자 BD	남	미혼	25	포루투칼어	브라질	UPS	-	-	게임	-	NO	NO
...	...	...	...	...	...	...	...	...	...	...	...	...
사용자 BE	남	미혼	19	포루투칼어	브라질	GJ	-	웹디자인	게임	-	NO	NO
											평균지수	0.21

(그림8) 샘플링 된 웹 사용자 소셜정보

즉, 질의패턴과 소셜지수는 상관관계가 있으며, 유사한 질의를 사용하는 사람들의 평균 소셜지수는 그렇지 못한 집단들과의 평균 소셜지수 보다 높게 나타난다.

5. 결론 및 향후 연구

본 연구에서는 텍스트 마이닝 기법을 웹 사용자별 질의셋에 적용하여 카테고리별로 유사한 질의패턴을 가지는 사람들을 클러스터링 하였고, 이들 집단이 사용한 질의와 접근한 URL 정보를 활용하여 검색의 편의성을 향상할 수 있었다. 또한, 유사한 질의를 사용하는 사람들의 소셜지수 비교분석 결과, 질의 패턴이 유사한 웹 사용자끼리의 평균 소셜지수가 다른 카테고리에 존재하는 용어를 자주 사용하는 웹 사용자들과의 평균 소셜지수보다 높게 나타난다는 것을 분석하였다. 따라서, 특정 분야에 소셜지수가 높은 웹 사용자들이 자주 사용하는 질의 및 URL 정보를 추출하여 제공 한다면 관련분야에 대해서 보다 적합하고 유용한 정보를 획득할 수 있다.

본 연구에서는 특정 주제에 관심이 높은 웹 사용자 클러스터링을 단지 질의에 포함된 용어 빈도수만을 고려하여 수행 하였으나, 향후에는 질의뿐만 아니라 클릭한 URL 페이지의 상주시간(resident time)을 클러스터링 시 고려한다면 더 신뢰성이 높은 정보를 검색할 수 있다. 또한 다의어 및 동의어에 대한 효율적인 처리가 수반된다면 질의나 카테고리에 대하여 더 만족할만한 URL 정보를 제공받을 수 있을 것이다. 또한 질의 로그와 해당 웹 사용자의 실제 소셜정보를 사용하여 분석한다면 검색패턴과 소셜지수 간의 관련성을 좀 더 정확하게 검증할 수 있을 것이다.

참고문헌

[1] Doug Downey, Susan Dumais, Dan Liebling, Eric Horvitz, Microsoft Research, “Understanding the Relationghip between Searchers′ Queries and Information Goals”, CIKM2008.

[2] In-Ho Kang, GilChang Kim, KAIST, “Query Type Classification for Web Document Retrieval”, SIGIR2003.

[3] Lars Backstrom, Dan Huttenlocher, Jon Kleinberg, Xiangyang Lan. Cotnell University, Ithaca NY. “Group Formation in Large Social networks: Membership, Growth, and Evolution”, KDD2006.

[4] 김용학, “사회 연결망 분석,” 전영사, 2003.

[5] Onix Text Retrieval Toolkit API Reference, http://www.lextek.com/manuals/onix/stopwords2.html

[6] Xiubo Geng, Tie-Yan Liu et. al., “Query Dependent Ranking Using K-Nearest Neighbor”, SIGIR2008.

[7] Barry Wellman, Jamet Salaff, Dimitrina Dimitrova, Laura Garton, Milena Gulia, Caroline Haythornthwaite. University of Toronto. “COMPUTER NETWORK AS SOCIAL NETWORKS: Collaborative Work, Telework, and Virtual Community”, 1996.

[8] STEVEN M. BEITZEL and ERIC C. JENSEN, ″Automatic Classification of Web Queries Using Very Large Unlabeled Query Logs″, ACM2007.