인지과학, 제19권 제4호
Korean Journal of Cognitive Science
2008, Vol. 19, No. 4, 459~476.

동시출현 자질과 집단 지성을 이용한 지식검색 문서 사용자 명성 평가*

이 현 우[1] 한 요 섭[2] 김 래 현[2] 차 정 원[1†]

[1]창원대학교 컴퓨터공학과 [2]한국과학기술연구원

많은 사용자들의 참여로 구축된 집단 지성을 이용한 지식 검색 서비스에서 사용자가 원하는 답변을 빨리 찾고자 하는 요구가 증가하고 있다. 기존의 연구에서 조회 수, 추천 수, 답변 수와 같은 비텍스트 정보가 답변을 평가하는데 좋은 자질임이 증명되었고, 신뢰도를 추정할 수 있는 여러 종류의 단어 사전을 이용하여 답변의 좋고 나쁨을 평가할 수 있는 연구도 진행되었다. 하지만, 조회 수, 추천 수, 답변 수와 같은 비텍스트 정보는 사용자 조작이 간단하여 지속적으로 관리를 해야 하며, 신뢰도를 추정할 수 있는 단어는 지속적으로 보강되어야 한다. 본 논문에서는 이러한 문제점을 해결하고자 동시출현 자질을 이용한 질문과 답변의 유사성을 활용하여 집단 지성에서 사용자의 활동을 분석하여 사용자의 명성을 평가하는 방법을 제안한다. 사용자의 명성을 계산할 수 있다면 조회 수와 추천 수가 많지 않은 답변의 신뢰도도 비교적 정확하게 추정할 수 있다. 이를 위해 우리는 PageRank 알고리즘을 수정하여 사용자 명성을 계산한다. 네이버 지식iN의 문서로 실험한 결과, 기존 정답 선택률을 보완할 수 있는 결과를 보였다.

주요어 : 사용자명성, 집단지성, 동시출현자질, 페이지랭크, 알고리듬

---

* 본 연구는 지식경제부 및 정보통신연구진흥원의 IT핵심기술개발사업의 일환으로 수행하였음(과제관리번호: 2008-S-024-01, Rich UCC 기술개발).

† 교신저자: 차정원, 창원대학교 컴퓨터공학과, 연구세부분야: 자연어처리, 정보검색
E-mail: jcha@changwon.ac.kr

서 론

최근 인터넷은 집단 지성을 이용한 다양한 서비스를 선보이고 있다. 대표적인 서비스로 인터넷 무료 백과사전인 '위키피디아(Wikipedia, http://wikipedia.org/)', 사용자가 직접 제작한 동영상을 등록하여 공유할 수 있는 '유튜브(YouTube, http://youtube.com/)', 소셜 네트워크 사이트인 '페이스북(FaceBook, http://facebook.com/)', 국내에서는 '네이버 지식iN(http://kin.naver.com/)', 'Daum 신지식(http://k.daum.net/)'과 같은 지식검색 서비스를 예로 들 수 있다.

위 서비스들의 공통점은 인터넷을 사용할 수 있는 수많은 사용자의 자발적인 참여로 생성된 컨텐츠를 기반으로 서비스를 한다는 점이다. 그래서 생성된 컨텐츠를 쉽게 검색하기 위해 자체적으로 개발한 검색엔진을 통하여 사용자가 좀 더 쉽게 컨텐츠를 찾을 수 있도록 도와준다.

하지만, 위의 서비스에서 제공하는 검색엔진을 사용하여 원하는 컨텐츠를 찾는 것은 사용자에게 많은 부담을 준다. 생성된 컨텐츠가 많을 뿐더러, 동일한 컨텐츠를 다수의 사용자가 등록하여 동일한 컨텐츠가 다수개 검색되는 경우가 많기 때문이다.

특히, 국내업체가 서비스하는 '네이버 지식iN'과 '다음 신지식'인 경우, 사용자가 등록한 질문에 대해 불특정 사용자가 답변을 등록하고, 질문을 등록한 사용자가 알맞은 답변을 선택하여, 다른 사용자들이 질문을 검색할 때, 비슷한 질문에 대해 선택된 답변을 찾기 쉽도록 서비스하고 있다. 하지만, 질문에 대한 답변이 선택되지 않았을 경우, 질문에 대한 답변을 찾을 때까지 사용자는 비슷한 질문에 대한 수많은 답변을 하나하나 열람하여 확인을 해야 한다는 문제점을 가지고 있다. 그래서 이러한 문제점을 해결하고자 지식검색 문서의 품질을 평가할 수 있는 다양한 연구가 진행되고 있지만, 아직 미흡한 편이다.

지식검색 문서 품질 평가의 초기 연구 단계에는 조회 수, 추천 수, 답변 수와 같은 지식검색 서비스 업체에서 측정할 수 있는 비텍스트 정보를 이용하여 문서의 품질을 평가하는 방법이 연구되었다[1].

본 논문에서는 사용자의 명성을 평가하여 새롭게 제시된 답변의 품질을 평가한다. 하지만 위와 같은 단순한 비텍스트 정보는 조작이 간편할 뿐더러, 먼저 작성된

문서보다 비텍스트 정보가 부족하여 나중에 작성된 문서가 아무리 높은 품질을 가지고 있다하여도 항상 낮은 순위를 가지므로 올바른 문서 품질 평가 불가능하다.

지식검색 문서의 질문과 답변은 매우 관련성이 높다. 대부분 질문에 나온 내용을 기반으로 사용자들이 답변을 작성하기 때문에 질문과 답변은 유사도가 높으며, 반대로 선택되지 않은 답변은 유사도가 많이 떨어진다. 그래서 질문과 답변의 유사도를 질문에 출현한 단어와 답변에 출현한 단어의 동시출현 자질 정보[2]를 이용하고, 사용자 명성이 높은 사람의 답변이 우수한 답변으로 추천될 수 있도록 함으로써 이러한 단점을 극복하려고 한다.

이후, 본 논문의 구성은 다음과 같다. 2장에서는 기존 연구에 대하여 논하고, 3장에서는 본 논문에서 제안하는 동시출현 자질, 그리고 집단 지성에서 사용자의 활동을 분석하는 '사용자 명성' 알고리즘을 설명한다. 4장에서는 본 연구 실제 지식검색 서비스 업체에서 제공하는 답변 선택률과 본 논문에서 추정된 사용자 명성을 비교하며, 마지막으로 5장에서는 결론 및 향후 연구에 대하여 언급한다.

기존 연구

수많은 문서에서 특정 질의와 관련성이 높은 문서는 검색할 수 있어도, 검색된 문서 중에 권위(authority) 있는 문서는 질의의 관련성만으로는 찾기가 힘들다.

하지만, 어떤 문서에 특정 문서로 향하는 하이퍼링크(hyperlink)는 문서를 작성자의 판단이 인코딩 되어 있는데, 중요한 문서일수록 그 문서로 향하는 하이퍼링크의 개수가 많아진다. 그래서 [3]은 권위 있는 문서를 특정 질의와 관련성 높은 문서들의 하이퍼링크의 구조를 분석하여 해결하고자 하였다.

[4]는 [3]처럼 하이퍼링크의 구조를 분석해서 문서의 중요도를 나타내고자 하였지만, [3]과 다르게 특정 문서로 향하는 하이퍼링크의 개수가 일반적인 의미의 중요성과 다를 수 있다고 하였다. 다수의 문서가 A라는 유명한 문서를 가리키고 있으며, B문서가 A문서에 하이퍼링크 되어 있을 경우, 유명한 A문서에서 B문서로 가는 하이퍼링크이므로 매우 중요한 링크라고 할 수 있다. 그래서 A문서에서 하이퍼링크한 B문서는 A문서를 하이퍼링크한 다수의 문서보다 더 높은 중요도를 가질

수 있도록 하는 알고리즘인 PageRank를 제안하였으며, 하이퍼링크의 구조만으로 검색될 문서의 중요성을 해결하였다.

북마크(Bookmark)를 등록하여 공유할 수 있는 'del.icio.us(http://del.icio.us/)'처럼 문서에 단어가 거의 존재하지 않으며, 문서를 대표하는 단어의 집합인 태그(tag, folksonomy)만 존재할 때, 단어를 이용한 검색은 출현하는 단어가 극히 제한적이라 검색에 많은 어려움이 있다. 그래서 [5]는 문서, 작성자, 태그를 이용하여 구축할 수 있는 방향성 있는 그래프(directed graph)에 [3]의 PageRank을 적용한 FolkRank를 제안하였다. 이는 유명한 문서를 대표하는 태그는 유명한 작성자가 작성한다는 가정에서 출발하였는데, 제안한 FolkRank에서 각 태그의 중요도를 계산하여, 검색 시 문서에 출현한 태그의 중요도를 이용하여 문서의 랭크를 계산하였다.

[3, 4, 5]처럼 문서들의 하이퍼링크의 구조만으로 문서를 신뢰도를 나타내고자 할 때, 새로 생성된 문서가 기존에 존재하는 문서들과 하이퍼링크가 아직 부족하므로 상위 순위에 위치한다는 것은 힘든 일이다. 그래서 [6, 7]은 문서의 중요도를 평가하기 위해서 문서의 내용을 평가할 수 있는 정보와 평가하는데 방해가 되는 잡음의 비율(information-to-noise ratio)을 사용하였다.

국내에서는 [6, 7]과 비슷하게 [8]에서 문서의 신뢰도를 측정하기 위해서 신뢰도를 추정할 수 있는 자질의 사전을 구축하여, 자질을 이용한 품질 평가 모델을 생성, 문서의 신뢰도 등급을 나누었다.

지식검색 문서의 작성자 신뢰도 평가

동시출현 자질 정보

지식검색 문서의 질문과 답변에서 유사도를 측정하고자 동시출현 자질 정보를 이용하고자 한다. 하지만 [2]에서 블로그의 스팸 덧글을 분류하고자 블로그 본문에 있는 주제어(명사류)와 덧글에 있는 주제어의 동시출현 자질 정보를 이용하였는데, 본 논문에서는 주제어를 사용하지 않고, n-gram을 이용한 동시출현 자질 정보를 사용하고자 한다. 이는 특정 분야에 사용하는 주제어인 경우, 축약형이 많으며 특

정 분야의 전문가들만 사용하는 단어를 많이 사용하므로, 범용으로 제작된 품사부착기 또는 명사추출기는 주제어를 찾기 힘들 뿐더러, 찾는다 하더라고 오류를 범할 수 있는 확률이 높아, 오히려 잘못된 주제어 정보를 제공할 수 있으므로 사용하지 않는다.

먼저, 질문의 제목과 본문에서 추출한 n-gram과 답변의 본문에 출현한 n-gram으로 동시 출현한 n-gram을 수집한다. 동시출현 자질을 이용하여 문서 X와 문서 Y의 유사도를 측정하기 위해 식 (1)을 사용한다.

$$s(X, Y) = \frac{|X \cap Y|}{|X| + |Y|}, \tag{1}$$

여기서 |X|는 문서 X의 n-gram수를 나타낸다. X는 질문, Y는 답변이 된다.

집단 지성을 활용한 사용자 명성 추정

집단 지성을 이용한 사용자 명성 추정은 [4]에서 하이퍼링크를 이용한 PageRank 알고리즘을 기반으로 작성되었다. PageRank에서는 단순히 문서와 문서사이의 링크 관계를 분석한다. 본 논문에서는 제안하는 '사용자 명성(User Reputation)' 알고리즘은 질문과 답변의 작성자 모두를 하나의 문서로 가정하고 질문과 답변의 작성자 사이에 가상의 링크를 생성하여 링크를 분석하는 단계까지는 PageRank와 비슷하지만 링크의 종류에 따라 가중치를 달리하여, 사용자의 명성을 추정한다는 점이 PageRank와 다른 점이다.

사용자 명성에서 사용하는 링크의 종류는 '답변으로 선택된 링크'와 '답변으로 선택되지 않은 링크', 총 2가지이다.

[그림 1]에서 사각형은 질문을 나타내며, 타원은 답변의 작성자를 나타내며, 사각형과 타원에 있는 숫자와 알파벳은 이들을 구별하기 위한 고유ID이다. 실선으로 된 화살표는 '답변으로 선택된 링크'를 나타내며 점선으로 된 화살표는 '답변으로 선택되지 않은 링크'를 나타낸다. 특정 답변을 작성한 사용자를 그 질문에서 사용자로의 링크로 표현한다.

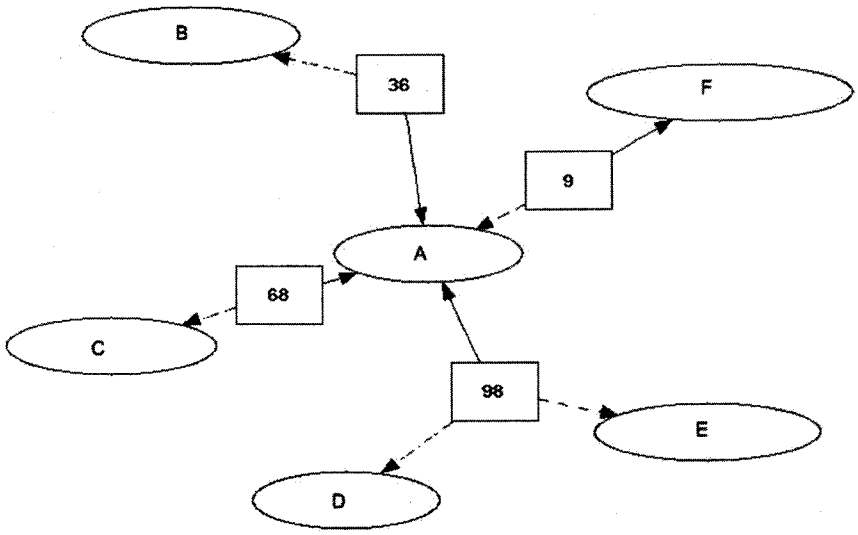

그림 1. 질문과 답변 작성자의 링크 예제. 사각형은 질문을 나타내고 타원은 답변 작성자를 나타낸다.

$$UR(p_i) = (1-d) + d \sum_{q_i \in M(p_i)} \frac{UR(q_i) \times f(q_i)}{C(q_i)}, \tag{2}$$

p_i : 답변 작성자

q_i : 질문

$UR(p_i)$: p_i의 *UserReputation*값,

$M(p_i)$: p_i가 답변을 작성한 모든 질문,

$C(q_i)$: q_i에 작성된 모든 답변의 수,

$f(q_i)$: $\begin{cases} q_i \text{에서 답변으로 선택되었을 경우} & : 0.8 \\ q_i \text{에서 답변으로 선택되지 않았을 경우} & : 0.2 \end{cases}$

$d = 0.85$

식 (2)는 사용자 명성(User Reputation)을 계산하는 식으로 위의 사용자 명성 값이 수렴할 때까지 반복 수행한다. $f(q_i)$는 q_i가 작성한 답변에 대해 가중치 역할을 하는

데, 이는 선택된 답변이 좀 더 높은 사용자 명성 값을 가지도록 하기 위함이다.

실 험

실험 데이터

본 연구에서 사용한 실험 데이터는 '네이버 지식iN'에서 수집한 자료로 특정 분야에 대해 답변 선택률이 높은 상위 100명의 사용자가 작성한 답변을 수집하였다.

표 1. 실험 데이터 정보

수집된 사용자	수집된 질문 개수	수집된 답변 개수	한 질문당 평균 답변 개수
20,900 명	20,588 개	43,913 개	2.13 개

동시출현 자질을 이용한 질문과 답변의 유사도 실험

동시출현 자질에 사용된 텍스트는 질문의 제목, 본문 그리고 답변의 본문이다. 먼저 n-gram을 이용한 동시출현 자질이므로 적당한 n-gram 크기를 선택해야 한다.

표 2. n-gram 크기에 따른 답변과 질문의 평균 유사도

n	선택된 답변과 질문의 유사도	선택되지 않은 답변과 질문의 유사도
2	0.676401251027	0.408916806536
3	0.408059686328	0.263355882131
4	0.181775941125	0.151945087754

단순히 n-gram의 동시출현 자질을 이용하여 계산된 유사도이지만, 질문의 내용을 바탕으로 답변이 작성된다는 지식검색 문서의 특징을 반영하고 있다. 특히 n이 2일 때의 '선택된 답변과 질문의 유사도'는 다른 n값에 비해 상당히 높은 값을 가

지고 있다. 그래서 본 논문에서 사용할 n-gram은 bi-gram이다.

집단 지성을 활용한 사용자 명성 추정 실험

[그림 2]는 실험 데이터에 식 (2)를 적용한 결과이다. 일반적으로 선택된 답변을 많이 작성한 사람이 높은 사용자 명성을 가지고 있다. 'U6'은 작성한 답변의 수는 사용자 명성이 높은 상위 25명의 사용자 중에 가장 높다. 하지만 선택된 답변이 적어 사용자 명성이 6번째로 높다. 'U3'은 작성된 답변의 개수는 적으나 선택된 답변의 개수가 주위 사용자들보다 많아 사용자 명성이 3번째로 높다. 'U6'이 작성한 답변보다 'U3'이 작성한 답변을 좀 더 높은 순위를 가질 수 있다면, 사용자가 '네이버 지식iN'에서 원하는 답변을 좀 더 빠르게 찾을 수 있을 것이다.

식 (3)은 식 (2)에서 선택된 답변과 선택되지 않은 답변의 가중치를 달리 했음에도 불구하고 전체적으로 작성한 답변이 많을 경우 높은 사용자 명성 값을 가지는 문제를 개선한 식이다. 식 (3)은 식 (1)의 n이 2인 bi-gram 동시 출현 자질을 이용하여 질문과 답변의 유사도를 적용하였다.

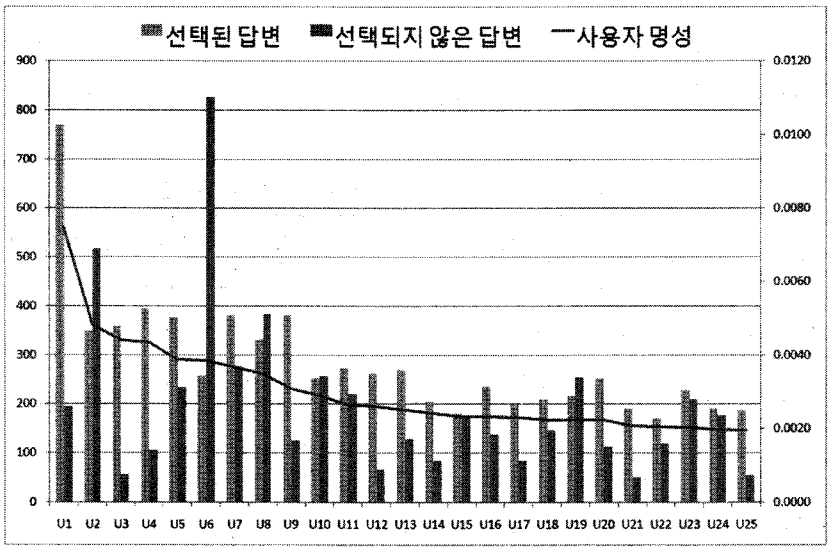

그림 2. 식 (2)를 적용한 사용자 명성, 가로축은 사용자

표 3. 식 (2)를 적용한 사용자 명성, 순위는 사용자 명성의 순위, ID는 사용자, UR은 사용자 명성, SR은 선택된 답변, NR은 선택되지 않은 답변, SUM은 작성한 답변 수, 사용자 명성 내림차순 정렬

순위	ID	UR	SR	NR	SUM
1	U1	0.0075	771	194	965
2	U2	0.0048	350	518	868
3	U3	0.0044	361	57	418
4	U4	0.0043	397	107	504
5	U5	0.0039	378	234	612
6	U6	0.0039	258	828	1,086
7	U7	0.0037	382	278	660
8	U8	0.0035	332	384	716
9	U9	0.0031	382	125	507
10	U10	0.0029	254	259	513
11	U11	0.0027	274	221	495
12	U12	0.0026	264	67	331
13	U13	0.0025	271	130	401
14	U14	0.0024	205	85	290
15	U15	0.0023	181	177	358
16	U16	0.0023	237	138	375
17	U17	0.0023	204	85	289
18	U18	0.0022	211	146	357
19	U19	0.0022	217	255	472
20	U20	0.0022	253	114	367
21	U21	0.0021	192	52	244
22	U22	0.0021	170	120	290
23	U23	0.0020	229	210	439
24	U24	0.0020	191	178	369
25	U25	0.0020	188	55	243

$$UR(p_i) = (1-d) + d \sum_{q_i \in M(p_i)} \frac{UR(q_i) \times f(q_i) \times s(q_i)}{C(q_i)}, \tag{3}$$

$s(q_i)$: 질문과 답변 사이의 유사도 = 식 (1)

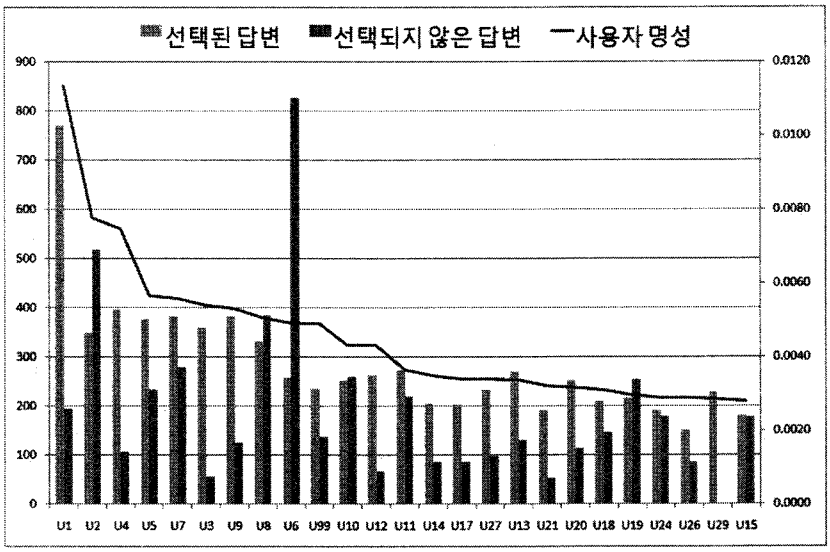

그림 3. 식 (3)을 적용한 사용자 명성, 가로축은 사용자, 왼쪽 세로축은 개수, 오른쪽 세로축은 사용자 명성 값

[그림 3]은 식 (3)을 적용한 실험이다. [표 3]과 [표 4]를 비교하면 전체적으로 사용자 명성 값이 많이 변했다는 것을 알 수 있다. 특히 'U6'경우 식 (2)를 사용했을 때보다 사용자 명성은 0.001 증가하였지만 전체적인 순위에서는 3위나 떨어졌다. 이는 아무리 많은 답변을 작성하였다고 해서 사용자 명성을 높일 수 없음을 알려주는 결과이다. 즉, 질문의 내용을 바탕으로 올바른 답변을 작성해야 높은 사용자 명성을 얻을 수 있으며, 이는 기존의 답변 수와 같은 비텍스트 정보를 이용하는 것 보다 더 좋은 성능을 보여줄 수 있는 자질임을 알 수 있다. 하지만 'U3'도 'U6'과 마찬가지로 순위가 3위나 떨어졌다. 'U3'은 선택된 답변의 개수가 382개이고 선택되지 않은 답변의 개수가 다른 'U7'과 'U9'의 사이에 존재하는데, 식 (3)을 사용해도 'U3'과 같이 작성한 답변 중에 정답으로 선택된 답변이 많은 사용자에게

표 4. 식 (3)을 적용한 사용자 명성, 순위는 사용자 명성의 순위, ID는 사용자, UR은 사용자 명성, SR은 선택된 답변, NR은 선택되지 않은 답변, SUM은 작성한 답변 수, 사용자 명성 내림차순 정렬

순위	ID	UR	SR	NR	SUM
1	U1	0.0114	771	194	965
2	U2	0.0078	350	518	868
3	U4	0.0075	397	107	504
4	U5	0.0057	378	234	612
5	U7	0.0056	382	278	660
6	U3	0.0054	361	57	418
7	U9	0.0053	382	125	507
8	U8	0.0051	332	384	716
9	U6	0.0049	258	828	1,086
10	U99	0.0049	237	138	375
11	U10	0.0043	254	259	513
12	U12	0.0043	264	67	331
13	U11	0.0037	274	221	495
14	U14	0.0035	205	85	290
15	U17	0.0034	204	85	289
16	U27	0.0034	235	97	332
17	U13	0.0034	271	130	401
18	U21	0.0032	192	52	244
19	U20	0.0032	253	114	367
20	U18	0.0031	211	146	357
21	U19	0.0030	217	255	472
22	U24	0.0029	191	178	369
23	U26	0.0029	151	85	236
24	U29	0.0029	229	1	230
25	U15	0.0028	181	177	358

높은 사용자 명성 값을 부여할 수 없었다.

지식검색 문서에서 좋은 답변이 있을 경우, 해당 답변에 대해 추천할 수 있는 기능이 있다. [그림 4]는 [1]과 같이 추천 횟수의 비율을 식 (2)에 적용한 새로운 식 (4)의 실험 결과이다. 추천 비율은 해당 답변의 추천 수에 질문에 대한 답변 전체의 추천 수로 나눈 값이다. 선택된 답변의 추천 비율을 높이는 이유는 선택된 답변이 높은 사용자 명성 값을 가지게 하기 위함이다. 반면, 자신의 사용자 명성을 높이고자 자신의 질문에 자신이 직접 답변을 작성하는 사용자가 있는데, 이는 사용자 명성 추정에 나쁜 영향을 줄 수 있으므로, 추천을 받더라고 낮은 비율이 적용될 수 있도록 하였다.

$$UR(p_i) = (1-d) + d \sum_{q_i \in M(p_i)} \frac{UR(q_i) \times f(q_i) \times s(q_i) \times r(q_i)}{C(q_i)}, \tag{4}$$

$$r(q_i): \begin{cases} \text{선택된 답변} & : \text{추천 비율} \times 0.6 \\ \text{선택되지 않은 답변} & : \text{추천 비율} \times 0.3 \\ \text{자문 자답인 답변} & : \text{추천 비율} \times 0.1 \end{cases}$$

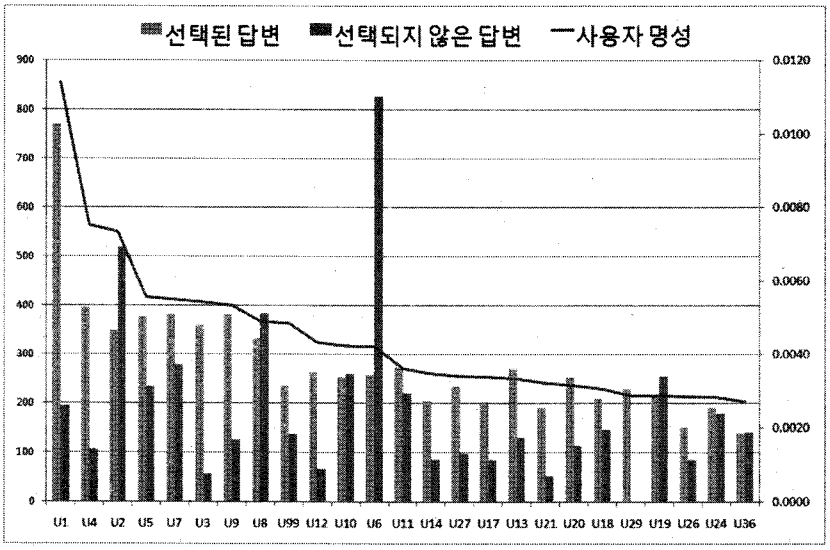

그림 4. 식 (4)를 적용한 사용자 명성, 가로축은 사용자, 왼쪽 세로축은 개수, 오른쪽 세로축은 사용자 명성 값

표 5. 식 (4)를 적용한 사용자 명성, 순위는 사용자 명성의 순위, ID는 사용자, UR은 사용자 명성, SR은 선택된 답변, NR은 선택되지 않은 답변, SUM은 작성한 답변 수, 사용자 명성 내림차순 정렬

순위	ID	UR	SR	NR	SUM
1	U1	0.0114	771	194	965
2	U4	0.0075	397	107	504
3	U2	0.0073	350	518	868
4	U5	0.0056	378	234	612
5	U7	0.0055	382	278	660
6	U3	0.0054	361	57	418
7	U9	0.0053	382	125	507
8	U8	0.0049	332	384	716
9	U99	0.0049	237	138	375
10	U12	0.0043	264	67	331
11	U10	0.0042	254	259	513
12	U6	0.0042	258	828	1,086
13	U11	0.0036	274	221	495
14	U14	0.0035	205	85	290
15	U27	0.0034	235	97	332
16	U17	0.0034	204	85	289
17	U13	0.0034	271	130	401
18	U21	0.0032	192	52	244
19	U20	0.0032	253	114	367
20	U18	0.0031	211	146	357
21	U29	0.0029	229	1	230
22	U19	0.0029	217	255	472
23	U26	0.0029	151	85	236
24	U24	0.0029	191	178	369
25	U36	0.0027	140	141	281

표 6. 자문자답한 사용자의 사용자 명성 순위가 낮아짐을 나타낸 결과

식 (3)			식 (4)		
순위	ID	자문자답	순위	ID	자문자답
1	U1	19	1	U1	19
2	U2*	18	2	U4	10
3	U4	10	3	U2*	18
4	U5	11	4	U5	11
5	U7	2	5	U7	2
6	U3	100	6	U3	100
7	U9	7	7	U9	7
8	U8	12	8	U8	12
9	U6*	12	9	U99	1
10	U99	1	10	U12	6

[표 4]와 [표 5]를 비교하면 추천 비율을 적용했지만 전체적으로 순위가 바뀌지는 않았다. 그 중에 'U2'와 'U6'의 사용자 명성 값에 변화가 있었다. 이는 보통 선택된 답변에 대해서만 사용자들이 추천을 하므로, 선택되지 않은 답변이 많은 사용자는 자연히 사용자 명성이 떨어지는 것을 확인할 수 있다. 하지만 [표 6]을 보면 이제까지 작성한 답변 중에서 선택된 답변의 비율이 높은 'U3'의 1/4에 해당하는 답변이 자문자답으로 작성된 답변으로 확인되었다. 즉 사용자들은 자문자답으로 작성된 답변에는 추천을 하지 않는다는 것으로 분석할 수 있다.

마지막으로 [그림 5]는 '네이버 지식iN'에서 제공하는 답변 선택률 순위[1]와 사용자 명성(식 4) 순위를 비교해 보았다. 전체적으로 일치하지는 않지만, 비슷한 결과를 보여주고 있다. '네이버 지식iN'의 특정 분야에서 1위한 사용자와 본 논문의 사용자 명성 1위의 사용자와 같은 사용자이다. 사용자 명성 그래프보다 '네이버 지식iN'의 그래프가 낮은 사용자는 'U3', 'U12'와 같이 작성한 답변에서 선택된 답변의 비율이 높은 사용자이다. 반대로 사용자 명성 그래프보다 '네이버 지식iN'의 답

1) 실제로 네이버 지식iN에서는 '답변 채택률'이라는 이름으로 제공된다.

변 그래프가 높은 사용자는 'U6'과 같이 작성한 답변에서 선택되지 않은 답변의 비율이 높은 사용자이다.

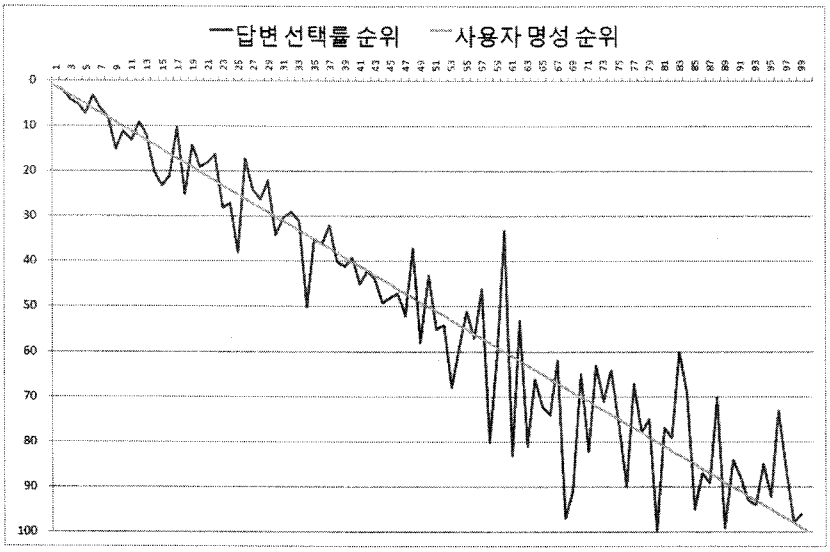

그림 5. 실제 답변 선택률 순위와 사용자 명성 순위, 가로축은 사용자, 세로축은 순위

결 론

본 논문에서는 지식검색과 같은 사용자 참여에 의해 작성된 질문과 답변을 이용하여, 사용자의 명성을 평가할 수 있는 방법을 제안하였다. 특히 동시출현 자질을 사용, 질문에서 선택된 답변과 선택되지 않은 답변의 유사도를 비교하여 아무리 많은 답변을 작성하였다고 해서 사용자 명성이 높아지지 않고, 질문의 내용을 바탕으로 작성된 답변 작성자의 사용자 명성을 높일 수 있는 자질을 개발하였다. 또한 기존의 추천 수, 댓글 수, 조회 수와 같은 비텍스트 정보를 사용할 경우 자료 부족 현상이 발생할 수 있으나, 사용자 명성은 사용자의 활동을 추적하여 평가하므로 사용자 명성이 높은 사람의 답변이 우수한 답변으로 추천되도록 할 수 있다.

실제 '네이버 지식iN'에서 수집한 실험 데이터로 사용자의 명성을 평가하여 '네이버 지식iN'에서 제공하는 사용자 답변 선택률 순위와 비교하여 비슷한 결과를 보였다. 모든 실험에서 사용된 데이터는 사람의 손을 거치지 않고, 있는 그대로 사용했다는 점에서 사용자 명성을 필요로 하는 다양한 인터넷 서비스에 적용될 수 있다는 가능성을 보였다.

또한, 동시출현 자질을 이용한 질문과 답변의 유사도를 평가할 때, 답변의 평가를 떨어뜨리는 동시출현 자질을 이용하여 유사도를 계산한다면 잡음의 비율을 줄일 수 있어서 사용자의 명성을 평가할 수 있는 더 좋은 자질이 될 수 있다. 그리고 [표 6]에서 자문자답에 대한 다양한 분석을 하여 사용자 명성에 반영하여 자문자답이 사용자 명성을 계산하는데 좋은 자질이 될 수 있도록 개발되어야 한다. 향후 연구 계획은 사용자 명성을 기반으로 실제 검색 모델에 적용하여 검색 성능을 효과적으로 향상시킬 수 있는 방법에 대해 연구하고자 한다.

참고문헌

[1] J. Jeon, W.B. Croft, J.H. Lee, and S. Park, "A Framework to Predict the Quality of Answers with Non-Textual Features". In Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Seattle, Washington, USA, 228-235, 2006.

[2] 전희원, 임해창, "본문과 덧글의 동시 출현 자질을 이용한 역카이제곱 기반 블로그 덧글 스팸 필터 시스템", 제19회 한글 및 한국어 정보처리 학술대회 (HCLT 2007) 논문집, 122-127, 2007.

[3] J. Kleinberg, "Authoritative Sources in a Hyperlinked Environment.", Journal of the ACM 46-5, 604-632, 1999.

[4] S. Brin, L. Page, "The anatomy of a large-scale hypertextual Web search engine.", Computer Networks and ISDN Systems 30, 107-117, 1998.

[5] A. Hotho, R. Jaschke, C. Schmitz, and G. Stumme, "Information retrieval in folksonomies: Search and ranking.", The Semantic Web: Research and Applications

4011, 411-42, 2006.

[6] X. Zhu, and S. Gauch, "Incorporating quality metrics in centralized/distributed information retrieval on the World Wide Web.", Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval, 288-295, 2000.

[7] Y. Zhou, and W.B. Croft, "Document quality models for web ad hoc retrieval.", Proceedings of the 14th ACM international conference on Information and knowledge management, 331-332, 2005.

[8] 이정태, 송영인, 임해창, "신뢰도 자질을 이용한 지식검색 문서의 품질 평가", 제19회 한글 및 한국어 정보처리 학술대회 (HCLT 2007) 논문집, 62-67, 2007.

1 차원고접수 : 2008. 12. 10
2 차원고접수 : 2008. 12. 19
최종게재승인 : 2008. 12. 22

인지과학, 제19권 제4호

(*Abstract*)

User Reputation Evaluation Using Co-occurrence Feature and Collective Intelligence

Hyun-woo Lee* Yo-Sub Han** LaeHyun Kim** Jeong-Won Cha*

*Dept. of Computer Engineering, Changwon National University

**Korea Institute of Science and Technology(KIST)

The user needs to find the answer to your question is growing fast at the service using collective intelligent knowledge. In the previous researches, it was proven that the non-text information like view counting, referrer number, and number of answer is good in evaluating answers. There were also many works about evaluating answers using the various kinds of word dictionaries. In this work, we propose new method to evaluate answers to question effectively using user reputation that estimated by the social activity. We use a modified PageRank algorithm for estimating user reputation. We also use the similarity between question and answer. From the result of experiment in the Naver GisikiN corpus, we can see that the proposed method gives meaningful performance to complement the answer selection rate.

Keywords : user reputation, collective intelligence, co-occurrence feature, pagerank, algorithm