TextRank를 이용한 키워드 정련

TextRank를 이용한 집단 지성에서 생성된 콘텐츠의 키워드 정련

Keywords Refinement using TextRank Algorithm

이현우, Hyun-Woo Lee*, 한요섭, Yo-Sub Han**,
김래현, LaeHyun Kim***, 차정원, Jeong-Won Cha****

요약 태그는 콘텐츠를 대표하는 신뢰도가 높은 키워드이다. 하지만 일부 기업과 사람들이 콘텐츠와 관련이 없는 키워드를 태그로 사용하여 본 논문에서는 무분별하게 사용된 키워드를 정련하는 알고리듬을 제안한다. 키워드 정련과 관련된 연구는 진행되지 않았지만, 본 논문에서는 단어와 단어사이에 가상의 링크를 생성, TextRank 알고리듬을 적용하여 콘텐츠에서 단어의 중요도를 계산하여 중요도가 낮은 단어의 일부를 콘텐츠의 제작자가 작성한 키워드에서 제거하여 키워드 정련을 하였다. 그 결과, 단순히 단어의 중요도가 낮은 하위 n%의 단어를 제거하는 방법보다는 신뢰도 구간을 만족할 때까지 제거하는 방법이 훨씬 좋은 키워드 정련 결과를 보였다.

Abstract Tag is important to retrieve and classify contents. However, someone uses so many unrelated tags with contents for the high ranking In this work, we propose tag refinement algorithm using TextRank. We calculate the importance of keywords occurred a title, description, tag, and comments. We refine tags removing unrelated keywords from user generated tags. From the results of experiments, we can see that proposed method is useful for refining tags.

핵심어: keywords, refinement, collective intelligence, textrank, algorithm

본 연구는 지식경제부 및 정보통신연구진흥원의 IT핵심기술개발사업의 일환으로 수행하였음. [과제관리번호: 2008-S-204-01, RichUCC 기술개발]

*주저자 : 창원대학교 컴퓨터공학과 e-mail: ggamsso@changwon.ac.kr
**공동저자 : 한국과학기술연구원 e-mail: emmous@kist.re.kr
***공동저자 : 한국과학기술연구원 e-mail: leahyunk@kist.re.kr
****교신저자 : 창원대학교 컴퓨터공학과 교수; e-mail: jcha@changwon.ac.kr

1. 서론

블로그(Blog), 위키위키(WikiWiki)와 같은 1인 미디어 시대를 대표하는 매체에서 자신이 작성한 콘텐츠를 대표하는 핵심어(키워드)를 태그(tags)라고 한다.

다음(Daum, http:/www.daum.net/)의 도움말에서는 아래 [그림 1]과 같이 태그를 정의하고 있다.

> "태그"는 자신이 쓴 글의 내용과 관련이 있는 단어입니다. 핵심어(키워드)라고 할 수도 있는데, 글을 쓸 때 자유롭게 태그를 입력해 두면 분류나 정리에 좋습니다. 또, 다른 사람들이 같은 태그로 어떤 글을 썼는지도 볼 수 있기 때문에 같은 관심사의 글을 찾기 쉽습니다.

그림 1, 태그(tags)의 정의

태그는 시스템에서 자동으로 생성되지 않으며, 사용자가 직접 입력하는 구조로 되어 있다. 그래서 해당 콘텐츠를 대표하는 단어로 높은 신뢰도를 가지고 있다. 이러한 태그의 특징을 이용하여 검색, 분류에 많이 사용되고 있다.

하지만, 콘텐츠의 핵심어를 찾는 훈련이 되지 않은 다수의 사용자들이 무분별하게 태그를 입력하여, 태그의 본질을 흐리게 하고 있다. 또한, 태그의 특성을 악용하여, 업체 또는 사용자가 자신이 작성한 콘텐츠와 관련 없는 다양한 종류의 태그를 입력하여 검색될 확률을 높이는 좋지 않은 방향으로 사용하고 있다.

본 논문에서는 콘텐츠를 대표하는 핵심어인 태그가 올바른 방향으로 사용될 수 있도록 TextRank 알고리듬[5]을 이용한 키워드 정련 알고리즘을 제안한다.

이후의 구성은 다음과 같다. 2장에서는 관련 연구에 대해 논하며, 3장에서는 본 논문에서 제안하는 키워드 정련 알고리즘을 설명, 4장에서는 실험 말뭉치에 대한 설명과 함께 키워드 정련 실험을 한다. 마지막 5장에서는 결론과 향후 연구에 대해 언급한다.

2. 관련 연구

키워드를 추천하는 연구는 다수 존재하나, 존재하는 키워드를 정련하는 분야의 연구는 아직 이루어지지 않았다. 그래서 정련에 대한 이전 연구 대신에 태그 추천 시스템과 관련된 연구와 문서의 중요도를 계산할 수 있는 연구로 대신하고자 한다.

[2]와 [3]에서는 사용자가 미리 입력해둔 태그에 추가할 수 있는 태그를 추천하는 시스템이다. 사용자가 콘텐츠에 입력한 태그와 콘텐츠에서 동시출현(co-occurrence)한 단어를 추출하여 후보 태그를 생성하여, 후보 태그를 평가하여 가장 점수가 높은 태그 중의 일부를 사용자에게 추천한다.

어떤 문서에 특정 문서로 향하는 하이퍼링크(hyperlink)는 문서를 작성자의 판단이 인코딩 되어 있는데, 중요한 문서일수록 그 문서로 향하는 하이퍼링크 개수가 많아진다. 그래서 [4]는 권위 있는 문서를 특정 질의와 관련성 높은 문서들의 하이퍼링크의 구조를 분석하여 해결하고자 하였다

[1]은 [4]처럼 하이퍼링크의 구조를 분석해서 문서의 중요도를 나타내고자 하였지만, [1]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 링크한 B문서는 A문서를 링크한 다수의 문서보다 더 높은 중요도를 가질 수 있도록 하는 알고리듬인 PageRank 알고리듬을 제안하였으며, 하이퍼링크의 구조만으로 검색될 문서의 중요성을 해결하였다.

그리고 문서의 내용을 요약하고자 중요한 문장을 계산하는 연구도 진행되었다. [5]에서는 문서에서 하나의 문장을 하나의 문서로 가정하여 문장과 문장사이에 유사도가 임계값 이상일 때 문장과 문장 간에 링크를 생성하여 [1]의 PageRank 알고리듬을 변형한 TextRank 알고리듬으로 문장의 중요도를 계산하여 중요도가 가장 높은 문장을 문서를 요약한 문장으로 사용하였다. 문장과 문장의 링크 방향에 따라 다른 성능을 나타내었지만 언어처리 도구를 사용하지 않아 다양한 언어에 적용할 수 있다는 특징을 가지고 있다.

3. TextRank 알고리듬을 이용한 키워드 정련

단어를 하나의 Text로 가정하고 단어와 단어사이에 가상의 링크를 생성할 수 있다면, 키워드 정련에 Text의 중요도를 계산하는데 사용하는 TextRank 알고리듬을 적용할 수 있다.

3.1 TextRank 알고리듬

TextRank 알고리듬은 Text A에서 Text B로 연결된 링크 하나를 Text A가 Text B에게 던지는 한 표로 해석하여 특정 Text의 득표수를 기준으로 중요도를 평가한다. 그리고 Text의 TextRank 값은 특정 Text의 중요도를 고려하여 중요한 Text로부터 표를 받은 경우 링크된 Text에 더 큰 TextRank 값을 부여한다(식1).

$$TR(p_i) = (1-d) + d \sum_{p_j \in M(p_i)} \frac{TR(p_j)}{L(p_j)} \qquad (1)$$

여기서 $TR(p_i)$와 $TR(p_j)$는 특정 Text의 TextRank 값이며, $M(p_i)$는 Text p_i가 링크하고 있는 모든 Text의 집합, $L(p_j)$는 p_j를 링크하고 있는 Text의 개수를 설명한다.

3.2 키워드 정련 알고리듬

콘텐츠에 댓글을 작성할 때는 콘텐츠의 제목, 본문, 키워드, 작성중인 바로 위의 댓글을 내용을 참조하여 작성되는 특징을 가지고 있다. 그래서 콘텐츠의 제목, 본문, 키워드에 출현한 단어는 댓글에 출현할 확률이 높다.

본 논문에서 제안하는 키워드 정련 알고리듬은 콘텐츠의 제목, 본문, 키워드, 댓글에 출현한 단어들 사이에 가상의 링크를 생성, TextRank 알고리듬을 적용하여 각 단어의 중요도를 계산하였다. 단, 같은 영역(제목, 본문, 댓글)에 같은 단어가 출현하여도, 모든 다른 단어로 정의하여 같은 단어가 많이 같은 영역에서 많이 사용될 경우, 더욱 많은 링크가 생성되도록 하였다.

먼저 댓글은 작성된 시간을 기준으로 정렬하여, 댓글과 댓글에서 출현한 단어들 사이에 가상의 링크를 생성하면 [그림 2]와 같은 결과를 얻을 수 있으며, 댓글은 이전에 작성된 댓글을 참조하여 작성된다는 성격이 반영되어 있다.

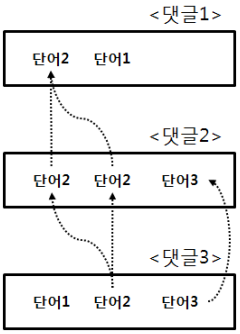

그림 2. 댓글과 댓글에 출현한 단어들의 링크

모든 댓글은 콘텐츠의 제목, 본문, 키워드를 참조하여 작성될 확률이 높으므로 [그림 3]과 같은 결과를 얻을 수 있다.

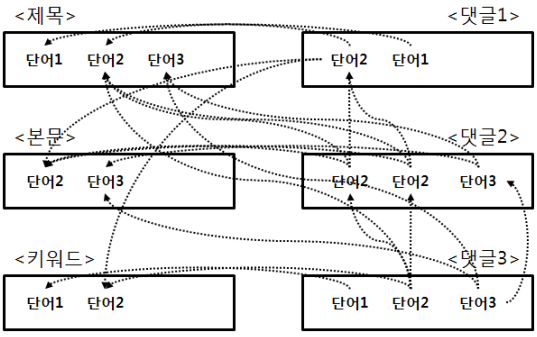

그림 3. 제목, 본문, 키워드, 댓글에 출현한 단어들의 링크

[그림 3]에서 링크가 생성된 모든 단어를 하나의 Text로 가정하여 (식1)에 적용을 하여, 단어들의 중요도인 TextRank 값을 계산한다.

4. 실험

4.1 실험 말뭉치

유투브(YouTube, http://www.youtube.com/)에서 'ipod'와 관련성이 높은 콘텐츠의 제목, 본문, 키워드, 댓글 모두가 존재하는 콘텐츠 80개를 수집하였으며, 키워드 정련 실험을 위해서 해당 분야의 전문가가 콘텐츠를 직접 보며 키워드를 작성하였다. 단, 콘텐츠의 제작자가 키워드에 사용한 단어와 해당 분야의 전문가가 키워드에 사용하는 단어가 다를 수 있다. 사전에 콘텐츠에서 어떤 단어가 출현했는지 전문가가 미리 확인하는 과정을 거쳤다.

표 1, 실험 말뭉치의 키워드 통계 정보, 단위(개)

	콘텐츠 제작자	해당분야 전문가
평균	9.2418	6.5443
분산	37.2568	6.7640
표준편차	6.1038	2.6007

4.2 실험 평가 기준

$$U' = U - r \tag{2}$$

$$P_k = \frac{|U' \cap G|}{|U'|} \tag{3}$$

$$R_k = \frac{|U' \cap G|}{|G|} \tag{4}$$

(2)의 U'는 정련된 키워드이며, U는 콘텐츠 제작자가 작성한 키워드, r은 U에서 제거해야할 키워드이다. 마지막으로 G는 해당 분야의 전문가가 작성한 키워드이다. (3)의 P_k는 정련된 키워드의 정확도이며 (4)의 R_k는 정련된 키워드의 재현률이다.

4.3 단어 간의 링크 생성 방법

TextRank 알고리듬에서 단어의 중요도를 계산할 때, 단어와 단어사이에서 링크를 생성하는 방법에 따라 다양한 결과를 가져올 수 있으므로, 본 논문에서는 [표 2]의 세 가지 방법으로 단어와 단어사이의 링크를 생성하였다. 단, 모든 단어에 대해서 링크를 생성하지 않고 불용어(stopword) 목록에 존재하지 않는 단어만 링크를 생성하였다.

표 2, 단어와 단어사이의 링크 생성 방법

링크 생성 방법	설명
①	단어의 중복을 허용하고 링크 생성
②	단어의 중복을 허용하지 않고 링크 생성
③	댓글간의 링크만 생성하지 않음

4.4 키워드 정련 실험

4.4.1 단어 중요도가 낮은 하위 n% 단어 제거 실험 – 실험 1

단어와 단어사이의 링크를 생성하여 TextRank 알고리듬을 적용하면, 실험 말뭉치에서 출현한 모든 단어들의 중요도를 계산할 수 있다. 정련된 키워드인 U'을 구하기 위해서 중요도가 낮은 단어를 콘텐츠 제작자가 작성한 키워드에서 제거를 해야 한다(1).

U는 콘텐츠 제작자가 작성한 키워드, r은 U에서 제거해야할 키워드이다. 본 실험에서는 콘텐츠에 출현한 모든 단어 중에 단어 중요도가 낮은 하위 n%에 해당하는 단어가 r에 속하며, 실험 결과는 [표 3]과 같다.

표 3, 단어 중요도가 낮은 하위 n%를 제거한 키워드 정련 실험

하위 n%	링크 생성 방법					
	①		②		③	
	P_k	R_k	P_k	R_k	P_k	R_k
99%	0.560	0.385	0.555	0.403	0.541	0.394
95%	0.641	0.668	0.652	0.695	0.626	0.752
90%	0.628	0.727	0.632	0.747	0.613	0.797
85%	0.612	0.787	0.605	0.793	0.612	0.801
80%	0.610	0.793	0.598	0.798	0.612	0.801
75%	0.612	0.801	0.599	0.805	0.612	0.801

단어 중요도가 낮은 하위 n%를 제거한 실험에서는 n의 값이 감소할수록 P_k와 R_k가 증가함을 알 수 있다. 이는 아무런 조건 없이 콘텐츠 작성자가 작성한 키워드를 제거하였으므로 의미 있는 키워드도 같이 제거될 수 있기 때문이다.

또한 콘텐츠 작성자가 작성한 키워드의 평균 개수는 약 9개이며 표준 편차가 약 6개이다. 이는 콘텐츠의 작성자가 최소 3개 이상의 키워드를 작성한다는 의미이며, 3개의 키워드가 존재하는 콘텐츠의 키워드는 정말로 중요한 단어를 사용하였을 경우가 높다. 이런 경우에도 아무런 조건 없이 키워드를 정련할 경우 중요한 키워드가 정련되는 문제점을 가질 수 있다.

그리고 링크 생성 방법에 따라 P_k와 R_k의 값이 바뀌는 것을 확인할 수 있는데, 특히 ①과 ②의 링크 생성 방법에서 차이를 보이고 있다. 이는 키워드에서 중요한 단어일수록 여러 부분에서 반복적으로 사용된다는 의미이며, 중요하지 않는 단어는 반복적으로 사용되지 않는다고 할 수 있다.

4.4.2 신뢰도 구간을 이용한 단어 제거 실험 – 실험 2

[실험 1]에서 아무런 조건 없이 콘텐츠 제작자가 작성한 키워드를 제거하여 콘텐츠의 제작자가 작성한 키워드에서 중요한 키워드까지 제거되는 문제점이 발생하였다.

이번 실험에서는 [실험 1]의 문제점을 해결하고자 신뢰도 구간을 설정하여 신뢰도 구간을 벗어나는 키워드만 제거하기로 하였으며, 실험 결과는 [표 4]와 같다.

표 4, 신뢰도 구간 90%일 때의 키워드 정련 실험

신뢰도 구간	링크 생성 방법					
	①		②		③	
	P_k	R_k	P_k	R_k	P_k	R_k
95%	0.636	0.878	0.638	0.872	0.634	0.875

신뢰도 구간을 적용함으로 P_k와 R_k가 [실험 1]의 P_k와 R_k보다 상당히 높아진 것을 알 수 있다. 불필요한 키워드가 많이 작성된 콘텐츠에서는 다량의 키워드가 정련되며, 콘텐츠에 작성된 키워드의 개수가 적으면 적을수록 소량의 키워드만 정련되는 결과를 가져왔다.

하지만 [실험 1]과 다르게 ①보다 ②의 링크 생성 방법이 더 높은 P_k와 R_k를 가지고 있다. 이는 다량의 키워드를 제거할 때는 적게 출현한 키워드를 제거하는 방법이 좋은 결과를 보였지만, 신뢰 구간을 만족시키는 부분까지 소량의 키워드를 제거할 때는 여러 분야에 골고루 출현하지 않은 단어를 제거하는 방법이 더 좋다고 해석할 수 있다.

아래의 [그림 4]는 실제 키워드 정련 결과를 나타낸 그림이다.

['~~1.1.3~~', '~~1.1.2~~', '~~apple~~', '~~no~~', 'jailbreak', 'ipod', 'apps', 'update', 'free', 'v1.1.3', '~~computer~~', 'iphone', 'hack', 'touch', 'installer', '~~automatic~~', 'tutorial']

['itouch', '~~mechanics~~', 'apple', 'gba', '~~viewer~~', 'jailbreak', '~~gadget~~', 'iradio', 'psx', 'nes', '~~game~~', '~~computer~~', 'video', 'iphone', 'ttr', '~~pdf~~', '~~electronics~~', 'ipod']

['~~h4ck3r~~', '~~apple~~', 'to', 'customize', '~~explanation~~', 'or', '~~an~~', 'trick', '~~how~~', 'iphone', 'install', '~~hack~~', 'touch', '~~ipot~~', '~~2.0~~', 'onto', 'ipod', 'tutorial', 'cool']

그림 4, 키워드 정련 결과, 괄호에 있는 단어는 콘텐츠의 키워드이며, 그 중에 취소선으로 표시된 단어는 본 논문의 키워드 정련 알고리듬으로 제거된 불필요한 키워드이다.

5. 결론

문서로부터 키워드를 추출하는 방법은 많이 연구가 되었지만, 작성되어 있는 키워드를 정련하는 연구는 아직 초기단

계이다. 아직까지 검증된 평가 방법이 존재하지 않으며, 실험에 사용할 말뭉치도 존재하지 않는다.

본 논문에서는 콘텐츠의 제목, 본문, 키워드, 댓글에 출현한 단어와 단어에서 가상의 링크를 생성하여 TextRank 알고리듬을 적용한 결과, 콘텐츠에서 출현한 단어의 중요도를 계산할 수 있었다. 계산된 단어의 중요도를 이용하여 콘텐츠의 작성자가 작성한 키워드에서 불필요한 키워드를 제거하여 키워드를 정련할 수 있었다.

불용어에 포함되지 않는 단어들만 링크를 생성하였으며 아무런 조건 없이 단어 중요도가 낮은 하위 n%의 단어를 제거하는 방법보다 신뢰도 구간을 설정하여 신뢰도 구간 밖의 단어만 제거하는 방법에서 더 나은 성능을 보였다.

다음 단계에서는 이렇게 정련된 키워드를 사용하여 실제 검색에서 정련 이전과 비교하여 우수한 결과를 나타내는 지에 대한 실험을 할 예정이다.

참고문헌

[1] S. Brin, and L. Page, "The anatomy of a large-scale hypertextual Web search engine," Computer Networks and ISDN Systems, vol. 30, no. 1-7, pp. 107-117, 1998.

[2] B. Sigurbjornsson, and R. van Zwol, "Flickr tag recommendation based on collective knowledge," 2008.

[3] R. Jaschke, L. Marinho, A. Hotho et al., "Tag Recommendations in Folksonomies," LECTURE NOTES IN COMPUTER SCIENCE, vol. 4702, pp. 506, 2007.

[4] J. Kleinberg, "Authoritative Sources in a Hyperlinked Environment," Journal of the ACM, vol. 46, no. 5, pp. 604-632, 1999.

[5] R. Mihalcea, and P. Tarau, "A language independent algorithm for single and multiple document summarization," Proceedings of IJCNLP2005, 2005.