웹의 개념지식을 위한 Anchor Text에서의 키워드 추출 알고리즘의 구현

조남덕,배환국,김기태
중앙대학교 컴퓨터공학과

A Implementation of Keyword Extraction Algorithm Using Anchor Text for Web's Conceptual Knowledge

Nam-Deok Cho, Hwan-Kook Bae, Ki-Tae Kim
Dept. of Computer Science & Engineering, Chung-Ang University

요 약

인터넷을 효과적으로 검색하기 위하여 검색엔진을 많이 이용하고 있다. 그런데 문서의 키워드를 추출할 적에 지금까지는 Anchor Text를 염두에 두지 않았었다. Anchor Text는 사람이 직접 요약한 것이고(요약성), 하이퍼링크를 포함하는 웹 문서에 반드시 존재하므로(보편성) 그 하이퍼링크가 가리키는 곳의 문서의 키워드를 추출에 적합한 용도가 될 수 있다. 웹 그래프는 이러한 Anchor Text를 이용하여 키워드를 추출함으로써 문서와 문서간, 단어와 단어간의 관계(연관성)까지도 나타내 줄 수 있게 한 검색 엔진 시스템이다. 그러나 Anchor Text 자체가 본문의 내용이 아니고, Anchor Text를 작성한 사람에 따라 다르게 작성되며, 본문의 내용과 무관한 내용도 작성할 수 있다. 따라서 Anchor Text 자체를 어떠한 여과 없이 문서의 키워드로 받아들이긴 힘들다. 본 논문에서는 TFIDF를 통해 좀 더 정확성이 있는 키워드를 추출하였다.

1. 서 론

인터넷은 정보의 바다라 할 만큼 많은 정보들이 존재하여 필요한 정보를 효과적으로 찾을 수 있는 방법이 필요하다. 현재 인터넷에서 정보를 찾기 위해서는 필수적으로 검색엔진이 사용되고 있다.

웹 그래프는 질의어를 입력하면 그와 연관되어 있는 단어들까지 보여주는 개념 기반 검색엔진이다.

그 연관어를 추출하기 위해 문서간의 연관성이 있는, 하이퍼링크의 anchor text를 이용하여 추출하였다. 그러나 본문 자체가 아닌 anchor text의 키워드 추출은 그 적합성을 검증 받지 못하였다.

본 논문에서는 그 적합성을 측정하며, 개선된 키워드 추출 알고리즘을 제시한다.

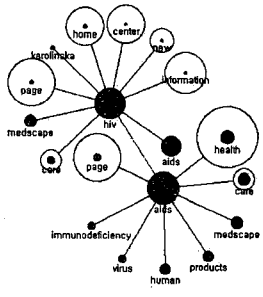
2. 기반 연구

2.1. 웹 그래프

웹 그래프란 웹 페이지간의 링크 정보를 이용하여, 사용자의 질의에 도식화된 개념 그래프를 보여줌으로써 원하는 정보를 쉽게 검색할 수 있도록 하는 검색엔진을 말한다[1].

웹 그래프는 질의어뿐만 아니라 그 질의어와 연관된 단어까지 보여줌으로써 사용자가 검색을 효과적으로 하게 해준다.

질의어는 aids 입니다.



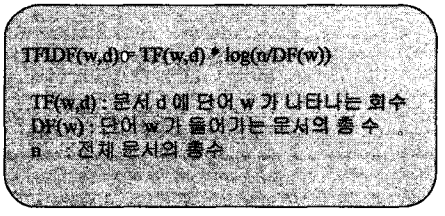
연관어는 질의어가 있는 문서에서 하이퍼링크로 연결된 문서의 핵심어로써 그 하이퍼링크의 Anchor Text에서 추출한 키워드이다.

2.1.1. 웹의 개념지식

하이퍼링크의 특징과 정보를 바탕으로, 요약 정보인 Anchor Text나 Alt Text에서 추출된 핵심어와 관련 웹 문서를 지칭하는 특성을 이용하여 구축된 개념그래프를 사용하여 구축된 지식을 웹의 개념지식이라 한다[1][2].

2.2. TFIDF

TFIDF 방식이란 하나의 문서 d에서 단어 w에 대한 weight값을 산출하는 방식으로 다음의 수식으로 표현할 수 있다.[3][4]



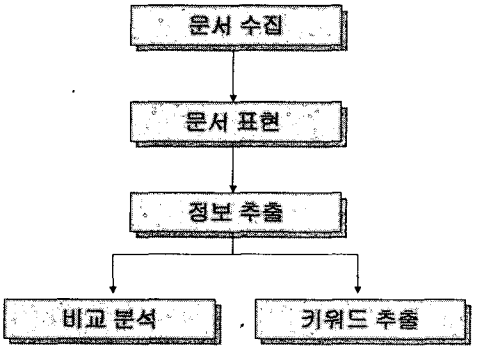
즉, 어떤 단어에 대한 중요도는 그 단어가 문서에 나온 횟수(Term Frequency)에 비례하고, 그 단어가 있는 모든 문서의 총 수(Document Frequency)에 반비례한다는 것이다.

TFIDF 방식을 이용하면 하나의 문서 중에서 가장 weight 값이 높은 단어가 그 문서에 키워드로 채택된다.

3. 시스템의 설계 및 구현

3.1. 시스템의 개요

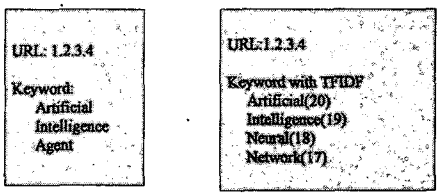
본 논문의 시스템의 전체적 구성은 밑의 [그림]과 같다. 문서 수집 단계에선 실험에 쓰일 웹 문서들을 모으는 단계이다. 그리고 문서 표현 단계에선 모아진 웹 문서들의 키워드를 TFIDF로 계산하기 위해 문서의 용어들을 최적화하는 단계이고, 정보 추출 단계에선 모아진 웹 문서들의 용어들간의 TFIDF 값을 계산하여 내림 차순으로 정렬하는 단계이다. 비교 분석 단계에선 TFIDF 값이 할당된 용어들과 웹 그래프의 hyperlink에서 추출된 키워드들과 비교 분석하는 단계를 말한다. 마지막으로 키워드 추출단계에선 Anchor Text에서 키워드를 추출하는 단계이다.



3.2. 비교 분석

각각의 URL마다 여러개의 키워드가 있을 수 있고 그 키워드를 위에서 계산된 키워드들과 비교를 한다. 현재 TFIDF로 계산된 키워드들은 내림 차순으로 정리하였으므로, 상위의 키워드들과 매치될수록 문서를 대표할 만한 키워드가 WebGraph에서도 추출되었음을 보여준다.

예를 들면 다음 그림과 같다.



하이퍼링크에서 추출된 키워드 실제 문서에서 TFIDF값을 갖는 단어들

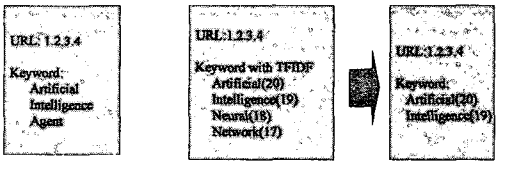
위의 그림에서 왼쪽의 것이 URL 1.2.3.4를 가리키는 하이퍼링크에서 추출된 키워드를 나타내는 것이고, 오른쪽의 것이 URL 1.2.3.4에 있는 실제 문서의 단어들(문서의 모든 단어들이며 TFIDF 값을 갖는다.)을 나타낸다.

위의 그림에서 볼 때, Artificial 이나 Intelligence 같은 경우, 실제 문서에서도 존재하며, 둘다 가장 높은 값을 가지므로 그 두 단어는 그 문서에 대해 중요한 단어라고 할 수 있다. 그러나 Agent 같은 경우 아예 문서에 존재하지 않으므로 중요하지 않은 단어라고 할 수 있다.

실제로는 문서에 나타나지 않아도 그 문서를 대표할 수 있는 단어가 있을 수도 있지만 여기서는 그런 경우는 없다고 가정한다.

3.3. 키워드 추출

Anchor Text에서 추출한 키워드가 반드시 본문에 나오라는 보장은 없다. 이러한 키워드는 제거하였다. 즉, 위 그림에서 하이퍼링크에서 추출한 키워드는 artificial, intelligence, agent이지만 agent는 실제적으로 문서에 존재하지도 않으므로, artificial 이나 intelligence만 키워드로 채택되는 것이다.



하이퍼링크에서 추출된 키워드 실제 문서에서 TFIDF값을 갖는 단어들 새로 추출된 키워드

위의 그림에서 보다시피 새로운 키워드 추출이 단순히 agent가 없어진 건만 뜻하는 게 아니라 TFIDF값도 그대로 갖고 있어 Ranking화가 가능하다는 것도 뜻한다.

4. 실험 및 평가

4.1. 비교 분석

tfidf(상위%)	10	20	30	40	50	60	70	80	90	100
추출된 키워드	31.2	40.3	54.2	59.8	64.3	66.9	71.2	73.4	75.9	77.6

결과에서 보면 TFIDF 값 상위 10%안에 Anchor Text에서 추출한 키워드 30%가 넘게 있음을 알 수가 있다. 그 30%의 키워드들은 문서를 대표할 수 있을만한 높은 수치를 갖는 단어라 할 수 있다.

그러나 상위 100%(단어 전체)에서 하이퍼링크에서 추출된 키워드와 매치되는 단어는 75%밖에 되지않음도 보여주고 있다. 이 것은 하이퍼링크의 키워드가 문서에 존재하지 않을 수 도 있다는 것을 보여주며, 그 키워드들은 문서에 필요없는 단어라 할 수 있다.

전체적으로 보면, Anchor Text에서 추출한 키워드가 문서상에 있으면, 대체적으로 높은 가중치를 가지고 있어서 키워드로써 적합한 반면에, 문서에 단 한번도 나오지 않는 키워드도 있어서 키워드를 추출할 때 이를 고려하여야 한다.

4.2. 키워드 추출

새롭게 추출된 키워드에 대한 평가는 기존의 웹 그래프에서 추출된 키워드와 비교해서 평가한다. 평가 기준은 IR 시스템의 평가기준인 정확율(Precision)과 재현율(Recall)로 하였다. 정확율과 재현율에 대한 식은 다음과 같다.

$$\text{정확율}(\text{Precision}) = \frac{\text{나온 문서 중에 질의에 관련있는 문서 수}}{\text{질의에 대해 나온 문서 수}}$$

$$\text{재현율}(\textit{Recall}) = \frac{\text{나온 문서중 질의에 관련있는 문서 수}}{\text{질의에 관련있는 문서의 총 수}}$$

여기에서 '관련 있는'의 기준은 문서에 그 질의어가 나오는 횟수로 하였다. 즉, 질의어가 문서에 몇 번 이상 나오는가를 임계치로 두어 그 임계치를 넘어서면 '관련 있는' 문서로 하여 실험을 하였다.

임계치	WebGraph		새로 제안한 모델	
	Precision	Recall	Precision	Recall
1	93.3	27.6	100	28.2
2	90.8	33.8	94.7	33.6
3	85.4	38.1	89.4	37.8
4	80.9	41.0	85.5	40.8
5	72.7	49.5	76.3	49.0
6	65.5	56.2	68.4	55.0
7	53.2	63.9	55.2	64.4
8	49.7	75.7	51.4	77.5
9	43.4	84.4	48.9	84.4
10	40.1	90.2	44.7	90.2

위의 표에서 보다시피 웹 그래프보다 새로 제안한 모델이 정확성이 더 높아졌다. 재현율은 전체적으로 비슷하게 나왔다.

4.3. Ranking

질의어에 대한 해당 문서를 찾아서 사용자에게 보여줄 적에 그 해당 문서들이 전부 관련이 있다고 하더라도, 관련 정도에 따라 보여주지 않으면, 해당문서가 너무 많아서 사용자가 전부 확인할 수 없을 적에, 별로 관련정도가 떨어지는 문서들만 볼 수있다. 그렇기 때문에 관련 정도가 큰 문서 순으로 보여주는 것이 좋다.

본 논문에서는 관련정도를 TFIDF 값으로 하여, 이를 웹그래프에 적용 시켰다.

5. 결론 및 향후 과제

본 논문에서는 어떤 웹 문서에서 키워드를 추출할 때, 본문 자체가 아닌 anchor text를 이용해서 추출할 때. 그 적합성을 측정하였으며, 그에 따라 좀 더 개선된 키워드 추출 알고리즘을 제시하였다.

이와 같은 방법은 정확성을 높여 주며, 웹의 개념지식을 구축하는데 더 효과적이다.

향후 과제로는 anchor text에서 키워드를 추출할 적에 TFIDF와 같이 본문에서 추출하는 방법이 아닌 anchor text만의 효과적인 방법을 착안해 내는 것이 필요하다 하겠다.

참고문헌

[1] 최준영, "인터넷상의 하이퍼링크를 이용한 개념 그래프 검색 시스템", 중앙대학교 석사 학위 논문, 1998

[2] 조민재, " 웹의 개념지식을 이용한 자동 시소러스 생성법의 설계 및 구현", 중앙대학교 석사 학위 논문, 1999

[3] Armstrong, R., Freitag, D., Joachims, T., Michell, t., "WebWatcher: A Learning Apprentice for the World Wide Web"' AAAI 1195 Spring Symposium on Information Gathering from Heterogeneous, Distributed Environments, Stanford, March 1995.

[4] Salton, G., Buckley, C., "Term-weighting approaches in automatic text retrieval," Information Processing and Management, 24(5), 513--523, 1988.

[5] William B.Frakes, Ricardo Baeza-Yates, Information Retrieval Data Structure & Algorithms, Prentice Hall Upper Saddle River, New Jersey 07458

[6] Salton, g., and M. Mcgill 1983. An Introduction to Modern Information Retrieval. New York:McGraw-Hill