김일성종합대학학보
정보과학

주체108(2019)년
제65권 제2호

웨브상에서의 패쎄지검색방법에 대한 연구

리청한, 김은하

위대한 령도자 **김정일**동지께서는 다음과 같이 교시하시였다.

《현시대는 과학과 기술의 시대이며 이르는 곳마다에서 요구하는것은 기술입니다. 기술을 몰라가지고서는 경제조직사업과 생산지휘를 바로할수 없으며 사회주의건설에 적극 이바지할수 없습니다.》(《**김정일**전집》 제2권 499～500페지)

웨브상에서 1개 문서는 보통 여러개의 독립적인 부분주제들로 이루어져있는데 이러한 문서를 다중주제문서라고 한다.

이와 같은 다중주제문서를 1개 주제에 대응하는 패쎄지들로 분할한 다음 패쎄지검색을 진행하여 검색된 패쎄지를 문서검색의 결과로 출력하면 문서검색의 성능을 개선할수 있다.

현재 리용되고있는 정보검색체계들은 대체로 전체 문서가 검색단위로 취해진다. 즉 검색대상이 문서들의 모임(collection)이며 검색결과 역시 순위화된 문서들의 모임이다.

이로부터 검색체계사용자들은 검색된 문서를 보면서 필요한 정보를 찾지 않으면 안된다. 이러한 리유로 하여 현재의 정보검색체계들은 다중주제문서검색에 적합하지 않다.

보통 웨브자료는 서로 다른 길이를 가지는 GB정도의 문서들로 구성되여있다. 긴 문서는 일반적으로 여러가지 서로 다른 내용들을 포함하는데 이것들의 매개는 부분주제에 대응되며 이때 매개 부분주제를 1개의 주제만을 가지는 짧은 문서로 볼수 있다.

선행연구[1, 2]에서는 웨브상에서의 검색에 1개 문서전체를 하나의 검색단위로 취하였다. 그러나 이 방법은 다음과 같은 리유로 하여 다중주제문서검색에 적합하지 않다.

우선 다중주제문서의 1개 부분이 질문에 적합할 때 다른 부분은 그 질문에 적합하지 않을수 있다. 이런것으로 하여 전체 문서와 질문사이의 류사성이 떨어질수 있다.

다음으로 어떤 문서가 질문에 전혀 부정확한 다중주제문서인 경우라고 할지라도 질문확장방법을 리용하면 사용자의 질문에 적합한 문서가 검색될수 있는 가능성이 있다.

론문에서는 이와 같은 문제를 해결하기 위하여 HTML구조에 내장되여있는 구조특징을 리용하여 웨브상에서의 다중주제문서를 패쎄지를 단위로 하는 단일주제문서로 분할한 다음 패쎄지검색을 진행하여 문서의 검색성능을 개선하는 방법을 제안하였다.

1. 패쎄지분할을 위한 특징선택

다중주제문서들은 매 주제에 따라 패쎄지들로 분할할수 있다. 전체 문서대신에 부분주제를 대표하는 패쎄지를 사용하는것에 의하여 문서검색의 성능을 개선할수 있다. 그런데 여기서 문제로 되는것은 문서를 주제에 따라 분할할 때 분할경계를 무엇으로 설정하는가 하는것이다. 주요한 리유는 다른 종류의 내용을 포함하는 문서들을 패쎄지들로 분할하는것이 어렵기때문이다.

그러나 웨브환경에서 다중주제문서를 패쎄지로 어떻게 분할하는가 하는것은 HTML 구조를 사용하여 해결할수 있다. 그것은 비록 HTML언어가 반구조화된 언어이지만 HTML문서안에 존재하는 특징들이 의미정보를 포함하고있기때문이다.

론문에서는 이와 같은 특징들이 무엇인가를 밝히고 그것들을 새로운 부분주제의 시작표식으로 사용한다.

문서들이 표식에 의하여 패쎄지로 명백히 분할되는것으로 하여 웨브자료우에서의 패쎄지검색을 진행할수 있다.

웨브자료우에서의 패쎄지검색과정은 다음과 같다.

① 웨브자료를 해석한 다음 THML구조특징을 리용하여 새로운 부분주제의 시작표식을 찾아낸다.

② 2개의 표식을 포함하는 부분을 1개 패쎄지로 결정한다.

③ 결정된 모든 패쎄지들을 색인작성하고 패쎄지검색을 진행한다.

④ 같은 문서에 있는 패쎄지들을 순위화하여 기초문서의 대표문서로 변환하여 문서목록을 귀환시킨다.

패쎄지구분을 위한 특징선택에는 독립적인 부분주제를 포함하는 보통 저자에 의하여 삽입된 어떤 이름붙은 표식이 있다.

전형적인 이름붙은 표식은 프락탈이란?이다.

그다음 하이퍼련결은 문서의 URL대신에 이름가진 표식을 지적할수 있다.

실례로

 프락탈이란?

혹은

 프락탈이란?

이다. 즉 하이퍼련결이 내부에 있으면 이름가진 표식은 같은 페지에 있게 된다.

일반적으로 다중주제문서는 시작에서 머리부목록을 가지고있다. 목록은 내부하이퍼련결의 내용으로서 이것들의 매개는 문서에서 명백히 부분주제를 암시한다. 결과 사용자들은 대응하는 하이퍼련결을 찰칵하여 부분주제를 선택할수 있다.

따라서 이름을 가진 표식자는 부분주제의 시작을 표식하기 위한 좋은 특징으로 된다.

이름가진 표식자를 만날 때마다 부분주제의 표식을 끼워넣으며 매 문서의 시작에 부분주제의 표식을 넣는다.

결과적으로 문서의 패쎄지번호는 그 안에 있는 부분주제표식의 번호와 같다.

매 문서에서 부분주제에 대응하는 패쎄지들은 그림 1과 같다.

결국 1개 문서에서 패쎄지의 개수는 부분주제의 개수와 같다.

우와 같은 방법으로 뽑은 모든 패쎄지들을 색인작성한 다음 패쎄지검색을 진행한다.

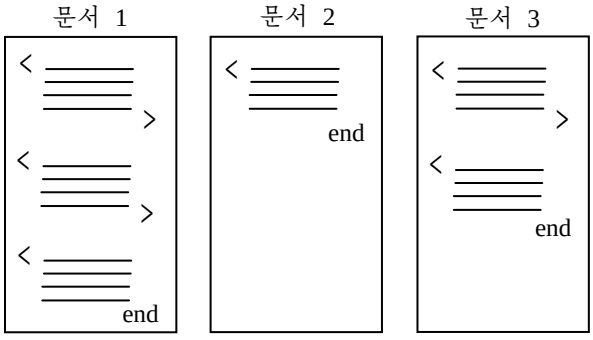

그림 1. 부분주제에 대응하는 패쎄지

2. 패쎄지검색

지금까지 패쎄지검색에는 tf-idf검색모형, BM25검색모형, 확률검색모형들이 많이 리용되고있다. 그러나 이 모형들의 본질적인 결함은 사용자질문이 짧은 경우 패쎄지의 순위화에 부정적인 영향을 준다는것이다.

이러한 결함을 극복하기 위하여 론문에서는 초기질문을 확장하여 패쎄지검색의 정확률을 높이는 방법을 제안하였다.

론문에서는 먼저 벡토르모형을 리용하여 초기질문을 가지고 패쎄지검색을 진행하여 패쎄지들을 순위화한다. 다음 순위화된 패쎄지에서 가장 웃준위패쎄지에 있는 용어들로 질문벡토르를 확장하고 이 질문벡토르를 리용하여 패쎄지검색을 다시 진행한다.

벡토르모형에서 패쎄지 p_j는 t차원벡토르로서 표현된다. 즉

$$\vec{p}_j = (\omega_{1j}, \omega_{2j}, \cdots, \omega_{tj})$$

이다. 여기서 ω_{ij}는 패쎄지 p_j에서 용어 t_i의 무게이다.

마찬가지로 질문벡토르 $\vec{q} = (\omega_{1q}, \omega_{2q}, \cdots, \omega_{tq})$이다. 여기서 ω_{iq}는 질문 q에서 질문용어 t_i의 무게이다.

벡토르모형에서는 질문 q에 대한 패쎄지 p_j의 류사성등급을 패쎄지벡토르 $\vec{p}_j$와 질문벡토르 $\vec{q}$사이 각의 코시누스로 평가한다. 즉

$$sim(\vec{p}_j, \vec{q}) = \frac{\vec{p}_j \cdot \vec{q}}{|\vec{p}_j| \times |\vec{q}|} = \frac{\sum_{i=1}^{t} \omega_{ij} \times \omega_{iq}}{\sqrt{\sum_{i=1}^{t} \omega_{ij}^2} \times \sqrt{\sum_{i=1}^{t} \omega_{iq}^2}}$$

이다. 여기서 $|\vec{p}_j|$와 $|\vec{q}|$는 각각 문서 및 질문벡토르의 노름이다.

패쎄지벡토르와 질문벡토르에서 용어무게는 여러가지 방법으로 계산될수 있지만 대체로 tf-idf방법을 리용한다. 즉

$$\omega_{ij} = f_{ij} \times \log\frac{N}{n_i}, \quad \omega_{iq} = \left(0.5 + \frac{0.5\, freq_{iq}}{\max_l freq_{lq}}\right) \times \log\frac{N}{n_i}$$

이다. 여기서

$$f_{ij} = \frac{freq_{ij}}{\max_l freq_{lj}}$$

이고 N은 패쎄지의 총개수, n_i는 질문용어 t_i가 나타나는 패쎄지의 개수이다.

$\omega_{ij\geq 0}$이고 $\omega_{iq\geq 0}$이므로 $sim(p_j, q) \in [0, 1]$이다.

이제 E를 질문확장을 위한 패쎄지벡토르모임이라고 하자. 즉

$$E = \left\{ p_j^+ \mid \frac{sim(p_j, q)}{\max_i sim(p_i, q)} \geq \tau \right\}$$

라고 하자. 여기서 q는 초기질문벡토르, τ는 류사성턱값이다.

E에서 패쎄지벡토르의 합 p_s를

$$p_s = \sum_{p_j^+ \in E} p_j^+$$

라고 하면 이것은 초기질문에 대하여 확장된 정보로 고찰할수 있다.

확장된 질문벡토르 q'는 다음과 같다.

$$q' = \frac{q}{\|q\|} + \alpha \frac{p_s}{\|p_s\|}$$

여기서 α는 무게조종을 위한 파라메터이다.

최종적으로 패쎄지들은 확장된 질문으로 류사도 $sim(p_j, q')$를 다시 계산하여 패쎄지들을 순위화한다.

매 문서에서 패쎄지들을 순위화한 다음 패쎄지검색결과를 문서수준의 결과로 변환한다. 즉 검색된 패쎄지가 어느 문서에 속해있는 패쎄지인가를 알아보고 그 기초문서의 대표문서로 패쎄지들을 출력한다.

매 문서에서 순위화된 패쎄지들은 그림 2와 같다.

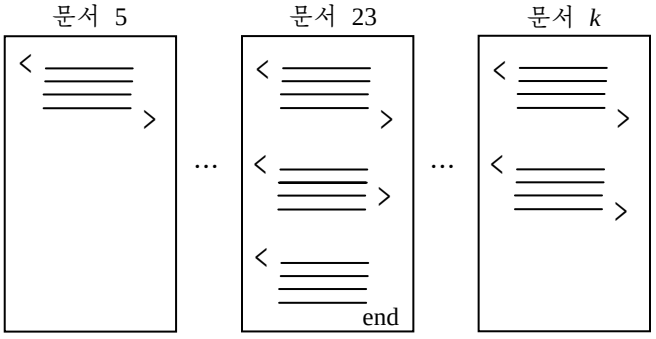

그림 2. 문서에서 순위화된 패쎄지

3. 실험결과와 분석

론문에서는 실험을 위하여 약 100만개의 웨브페지로 이루어진 크기가 약 10GB인 자료모임을 리용하였다.

표 1에 자료모임에서 다중주제문서량을 보여주었다. 표 1에서 보는바와 같이 문서의 약 11%가 다중주제문서이며 이것은 전체 자료모임의 약 37%크기를 차지한다.

표 1. 다중주제문서량

자료모임	총수	크기/GB
다중주제	188 947	2.66
전체	1 690 540	7.15

전통적인 문서검색과 패쎄지검색을 비교하기 위하여 두가지 실험을 진행하였다.

한가지 실험은 188 947개의 다중주제문서를 문서의 단위로 색인작성하고 50개의 질문

으로 꼭대기 1 000개의 문서를 얻었다.

다음 실험은 188 947개의 다중주제문서들을 HTML의 특징을 리용하여 모두 2 524 613개의 패쎄지로 분할한 다음 패쎄지검색을 진행하여 적합패쎄지를 포함하는 1 000개의 문서를 얻었다. 그리고 질문응답체계에서 패쎄지검색의 성능을 평가할 때 흔히 리용되는 평균정확률평가척도를 가지고 평가하였다.(표 2)

표 2. 검색성능평가결과

검색단위	평균정확률/%
문서	0.216 1
패쎄지(제안된 방법)	0.235 5(+8.98)

표 2에서 보는바와 같이 전체 문서검색보다 패쎄지에 의한 문서검색의 정확률이 훨씬 더 우월하다는것을 알수 있다.

다중주제문서에서 패쎄지에 의한 문서검색은 약 9%의 개선을 가져왔다. 이것은 패쎄지에 의한 문서검색이 다중주제문서속에서 매우 효과적이라는것을 보여준다.

맺 는 말

웨브상에서의 다중주제문서를 단일주제문서로 분할하는 방법을 HTML구조에 내장되여있는 구조특징을 리용하여 제안함으로써 문서의 검색성능을 개선하였다. 론문에서 제안한 방법은 웨브상에서 다중주제의 문서검색에서 그 효과가 뚜렷이 나타났다.

참 고 문 헌

[1] Rui-Hua Song et al.; Proceedings of the First International Conference on Machine Learning and Cybernetics, 9, 4, 2012.

[2] Neng Xie et al.; Proceedings of the First International Conference on Semantics, Knowledge, and Grid, 326, 2005.

주체108(2019)년 2월 5일 원고접수

The Study on Passage Retrieval Method on Web

Ri Chong Han, Kim Un Ha

In this paper using the feature of HTML structure we divide the multi-topic into only-topic document, and improve the retrieval performance of document using passage retrieval.

Key words: information retrieval, passage retrieval, HTML structure