퍼지를 이용한 스니핏 추출 방법

박선* · 최명수** · 김정호** · 김정욱** · 나희근** · 최석환** · 시우 쿠마** · 이성로***

*,**,***목포대학교

Snippet Extraction Method using Fuzzy

Sun Park* · Myeong Su Chio** · Cheong Ho Kim** · Cheong Uck Kim** · Hee Kun Na** · Seock Whan Choi** · Shiu Kumar** · Seong Ro Lee***

*,**,***Mokpo National University

E-mail : *sunpark@mokpo.ac.kr, ***srlee@mokpo.ac.kr

요 약

본 논문은 스니핏을 이용시 가끔 사용자의 의도와는 다른 잘못된 웹 페이지를 방문하는 문제를 해결하기 위해 퍼지를 이용한 새로운 스니핏 추출 방법을 제안한다제안방법은 의사연관 피드백을 이용하여 사용자의 질의를 확장학고, 확장된 질의와 웹 페이지 사이에 퍼지 연관을 이용함으로써 사용자의 의도가 의미적으로 더 잘 포함되는 스니핏을 추출할 수 있다

ABSTRACT

In order to solve problem which User sometime visits the wrong page with respect to user intention when uses snippet. this paper proposes a new snippet extraction method using fuzzy. The proposed method uses pseudo relevance feedback to expand the use's query. It uses the fuzzy association between the expanded query and the web pages to extract snippet to be well reflected semantic user's intention.

키워드

스니핏(snippet), 의사연관 피드백(pseudo relevance feedback), 퍼지연관(fuzzy association)

Ⅰ. 서 론

일반적으로 검색엔진이 제공하는 사이트의 추천 순서나 사이트 페이지의 요약 글은 사용자의 사이트 방문 여부에 큰 영향을 미친다. 검색엔진에서 보여주는 사이트 페이지의 요약 글을 스니핏(snippet)이라한다. 스니핏이란 웹 페이지가 나타내는 전체적인 내용을 포함하여서 대표할 수 있는 요약문을 의미한다. 요즘은 맞춤형 검색 등 개인화된 서비스가 증가하면서 개인화 스니핏에 대한 연구가 활발히 진행되고 있다

Huang은 스니핏 결과가 의미 있기 위해서는 사용자가 만족해야할 사항에 대하여 정의 하였다. 첫 번째는 사용자가 적은 노력으로 스니핏을 잘 구별 할 수 있도록 해야 한다. 두 번째는 사용자가 스니핏으로 부터 요점을 파악할 수 있도록 질의를 잘 표현해야한다[1].

본 논문은 Huang이 정의한 스니핏 목적을 만족할 수 있도록 의사연관 피드백과 퍼지를 이용한 새로운 스니핏 추출 방법을 제안한다. 연관 피드백의 종류는 질의 확장 시 사용자가 직접 개입하여 확장하는 연관 피드백과, 사용자의 개입 없이 자동으로 질의를 확장하는 의사연관 피드백이 있다[2, 3]. 퍼지는 퍼지집합 이론을 사용하여 정보검색 과정의 모호성을 정형화하는 방법으로, 문

장과 문장에 포함된 용어 간의 의미적 관계를 나타낼 수 있다[4, 5].

II. 본 론

본 논문에서 제안한 스니핏 추출 과정은 다음 같이 전처리, 의사연관 피드백, 스니핏 추출로 구성된다.

전처리단계에서는 검색 문서를 전처리하여 용어-문장 빈도행렬을 구성한다. 전처리 단계는 주어진 문서집합으로부터 불용어 제거, 어근추출, 용어빈도 벡터를 생성한다[6]. 불용어 제거는 Rijsbergen의 불용어 목록[6]을 이용하여서 목록에서 정의하고 있는 무의미한 용어들을 제거한다. 어근추출은 Porter의 어근추출 알고리즘[6]을 이용하여서 영어의 파생어들을 가장 중심이 되는 용어인 어근으로 변환한다.

의사연관 피드백 단계에서는 사용자의 질의를 확장한다. 본 논문에서는 일반적인 의사연관 피드백에 많이 사용하는 식(1)와 같은 양의 연과 피드백을 사용한다.

$$\overrightarrow{q^{new}} = \vec{q} + \sum_{\forall s_{*j} \in D_+} s_{*j} \qquad (1)$$

여기서, $\overrightarrow{q^{new}}$는 의사연관 피드백을 이용하여 새롭게 확장된 질의 벡터이고, $\vec{q}$는 사용자 질의 벡터이다. s_{*j}는 코사인유사도를 이용하여서 추출된 질의와 유사도가 가장 높은 문장 벡터이다.

스니핏 추출단계에서는 퍼지 연산자 값을 계산하여 확장된 질의와 포함 관계가 높은 문장을 추출하여 스니핏을 생성한다. 퍼지 포함관계 $\mu_{i,j}$가 최고 값을 가지면, d_i 문장을 j번째 문서의 스니핏 S_j 에 할당한다. 각각의 문장들이 각각의 스니핏 집합에 포함되는 정도인 퍼지 포함 관계[4, 5] $\mu_{i,j}$는 다음과 같이 정의 된다.

$$\mu_{i,j} = \sum_{\forall t_a \in d_i} \left[1 - \prod_{\forall t_b \in CT_j} (1 - r_{a,b}) \right] \qquad (2)$$

여기서, $\mu_{i,j}$ 는 j번째 문자의 스니핏 S_j에 i번째 문장 d_i가 속하는 정도이며, t_a는 a번재 용어를 t_b는 b번째 용어를 나타낸다. $r_{a,b}$는 용어 $t_a \in d_j$와 용어 $t_b \in CT_j$ 사이의 퍼지 관련 용어 관계이고, CT_j는 의사연관 피드백을 이용하여 확장된 질의에 포함된 j번째 용어 집합이다.

III. 결론

스니핏(snippet)이란 검색엔진이 사용자에게 제공하는 웹 페이지를 대표할 수 있는 요약된 정보로 사용자가 원하는 정보를 빠른 시간에 검색할 수 있도록 도와주는 역할을 한이다. 본 논문에서는 의사연관 피드백과 퍼지 연관을 이용한 새로운 스니핏 추출 방법을 제안하였다. 제안방법은 의사연관 피드백에 의해서 확장된 사용자의 질의와 웹 페이지 사이에 퍼지 연관 연산자를 이용함으로써 사용자의 의도가 의미적으로 더 잘 포함되는 스니핏을 추출할 수 있도록 하였다. 실험 결과 제안방법이 정보검색에 많이 사용하는 문장의 유사도에 의한 방법이나 의사연관 피드백을 기반으로 한 방법보다 더 좋은 성능을 보였다.

Acknowledgement

이 논문은 2011년도 정부(교육과학기술부)의 재원으로 한국연구재단의 대학중점연구소 지원사업으로 수행된 연구임(2011-0022980), 본 연구는 지식경제부 및 정보통신산업진흥원의 대학 IT연구센터 지원사업의 연구결과로 수행되었음 (NIPA-2012-H0301-12-2005)

참고문헌

[1] Y. Huang, Z. Liu, “Query Baised Snippet Generation in XML Search,” In proceeding of SIGMOD, pp.315-326, 2008.

[2] B. Y. Ricardo, R. N. Berthier, “Moden Information Retrieval,” ACM Press, 1999.

[3] S. Chakrabarti, “mining the web: Discovering Knowledge from Hypertext Data,” Morgan Kaufmann Publishers, 2003.

[4] C. Haruechaiyasak, M. L. Shyu, S. C. Chen, “Web Document Classification Based on Fuzzy Association”, In proceedings of the 25th Annual International Computer Software and Applications Conference (COMPSAC’02) (2002)

[5] L. A. Zadeh, “Fuzzy Sets, in Dubois, D., Prade, H. and Yager, R. R. editiors, Readings in Fuzzy Sets for Intelligent Systems”, Morgan Kaufmann Publiishers, 1993.

[6] W. B. Frankes, R. Baeza-Yaes, “Information Retrieval : Data Structure & Algorithms,” Prentice-Hall, 1992.