연관 피드백과 퍼지 함의 연산자를 이용한 스니핏 추출 방법

박선* · 심천식** · 이성로***

Snippet Extraction Method using Fuzzy Implication Operator and Relevance Feedback

Sun Park* · Chun Sik Shim** · Seong Ro Lee***

이 논문은 2011년도 정부(교육과학기술부)의 재원으로 한국연구재단의 대학중점연구소 지원사업으로 수행된 연구임 (2011-0022980), 이 논문은 2012년도 목포대학교 중형조선산업 지역혁신센터(RIC)에 의하여 지원되었음.

요 약

정보 검색 시 검색엔진은 사용자에게 웹페이지 순위와 웹페이지의 요약정보를 제공한다. 이중 웹 페이지를 대표할 수 있는 요약된 정보를 스니핏(snippet)이라한다. 스니핏은 사용자의 웹페이지 방문에 큰 영향을 준다. 정확한 방문 페이지의 정보를 모르고 단지 스니핏 만을 이용할 때에 가끔 사용자의 의도와는 다른 잘못된 웹 페이지를 방문할 수 있다. 이것은 검색엔진에서 지원하는 스니핏에 사용자의 의도를 정확하게 반영하는 것이 어렵기 때문이다. 본 논문은 이러한 문제를 해결하기 위해 연관 피드백과 퍼지 함의 연산자를 이용한 새로운 스니핏 추출 방법을 제안한다. 제안방법은 연관 피드백을 이용하여 사용자의 질의를 확장하고, 확장된 질의와 웹 페이지 사이에 퍼지 함의 연산자를 이용하여 질의와 확장된 질의의 포함관계가 반영된 스니핏을 추출함으로써 사용자의 의도를 스니핏에 더 잘 반영할 수 있다. 실험결과에서 제안방법이 다른 방법보다 스니핏 추출에 더 좋은 성능을 보인다.

ABSTRACT

In information retrieval, search engine provide the rank of web page and the summary of the web page information to user. Snippet is a summaries information of representing web pages. Visiting the web page by the user is affected by the snippet. User sometime visits the wrong page with respect to user intention when uses snippet. The snippet extraction method is difficult to accurate comprehending user intention. In order to solve above problem, this paper proposes a new snippet extraction method using fuzzy implication operator and relevance feedback. The proposed method uses relevance feedback to expand the use's query. The method uses the fuzzy implication operator between the expanded query and the web pages to extract snippet to be well reflected semantic user's intention. The experimental results demonstrate that the proposed method can achieve better snippet extraction performance than the other methods.

키워드

스니핏, 연관 피드백, 퍼지 함의 연산자, 퍼지 관계 곱

Key word

snippet, relevance feedback, fuzzy implication operator, fuzzy relational product

* 정회원 : 목포대학교 정보산업연구소 연구전임교수
(교신저자, sunpark@mokpo.ac.kr)
** 정회원: 목포대학교 조선공학과 조교수
*** 정회원 : 목포대학교 정보전자공학과 교수

접수일자 : 2011. 11. 09
심사완료일자 : 2011. 11. 16

Open Access http://dx.doi.org/10.6109/jkiice.2012.16.3.424

Ⅰ. 서 론

인터넷 상의 폭발적인 정보의 생성 및 증가는 사용자들에게 원하는 정보를 쉽게 찾을 수 있도록 검색엔진을 탄생시켰다. 특히 모바일 통신환경의 발전은 블로그, 온라인 뉴스와 트위터, 페이스북과 같은 소셜 네트워크 서비스(SNS, social network service) 등을 실시간으로 검색하도록 하면서 정보검색의 필요성을 더욱 증가 시키고 있다.

사용자가 검색엔진에 질의를 주었을 때, 검색엔진은 사용자에게 질의에 연관된 문서 목록의 제목과 함께 몇 개의 문장으로 구성된 요약문을 보여준다. 이들 중 요약문을 스니핏(snippet) 또는 엑셉트(excerpt)라 부른다. 스니핏은 사용자가 검색엔진에서 제시하는 모든 문서의 내용을 확인하지 않고도 필요로 하는 문서를 선택하는 데에 도와주는 역할을 한다.

이 때문에 일반적으로 검색엔진이 제공하는 사이트의 추천 순서나 사이트 페이지의 요약 글은 사용자의 사이트 방문 여부에 큰 영향을 미친다. 현재 대부분의 대형 검색엔진들은 질의 기반의 스니핏을 이용하고 있다[1].

현재 많이 연구되고 있는 스니핏 생성방법으로 Ko[2]와 Lin[3]는 통계적 모델에 기반을 둔 방법을 연구하였고, Penin[4]와 Huang[5] 등은 XML 기반의 방법을 연구하였으며, Turpin[6] 등은 압축방법을 이용한 스니핏 생성성능을 높이는 연구를 하였다. 또한 본 논문의 저자들은 이전에 용어들 간의 퍼지 상관관계를 이용한 스니핏 생성방법을 제안하였다[7]. 이들 방법은 용어의 출현빈도를 기반으로 하기 때문에 사용자의 의사를 스니핏에 정확히 반영 할 수 없을 때가 있다.

Huang[5]의 연구에서, 사용자가 적은 노력으로 스니핏을 잘 구별 할 수 있도록 해야 하며, 사용자가 스니핏으로 부터 요점을 파악할 수 있도록 질의를 잘 표현해야지 만 의미 있는 스니핏이 생성된다고 하였다[5]. 이는 현재의 질의기반의 검색엔진들이 질의에 따른 찾고자 하는 문서들의 조건을 충분히 만족하지 못하고 광범위한 의미로 확대되거나, 질의어가 동철이음이의어(heteronym)나 동음이의어(homonym)이어서 문서의 내용이 적절히 표현되지 못할 때 정확한 스니핏을 제공하지 못하는 단점을 가지고 있다.

본 논문은 Huang이 정의한 의미 있는 스니핏 을 생성 할 수 있도록 연관 피드백과 의미적 퍼지 함의 연산자를 이용한 새로운 스니핏 추출 방법을 제안한다. 제안방법은 질의를 잘 표현할 수 있도록 연관피드백을 이용하여 질의를 확장하며, 확장된 질의와 사용자 질의 간의 포함관계를 스니핏 추출에 이용함으로써 사용자의 의도가 의미적으로 더 잘 포함되는 스니핏을 추출한다.

연관 피드백은 질의를 확장하는 방법으로 질의와 연관된 문장을 질의에 더욱 근접하도록 질의를 수정하며, 질의와 연관이 없는 문장은 질의와 더욱 멀어지도록 질의를 수정한다. 연관 피드백의 종류는 질의 확장 시 사용자가 직접 개입하여 확장하는 연관 피드백과, 사용자의 개입 없이 자동으로 질의를 확장하는 의사연관 피드백이 있다[8, 9].

퍼지 함의 연산자(fuzzy implication operator)는 일반 함의 연산자를 확장하여 퍼지에 적용한 것으로 두 퍼지 집합에서 한 집합이 다른 집합에 포함되는 정도를 계산할 수 있다[10].

본 논문의 구성은 다음과 같다. 2장에서는 관련연구로 스니핏에 대한 관련연구 및 의사연관 피드백과 퍼지 함의 연산자에 대하여 알아본다. 3장은 제안방법으로 퍼지 연관을 이용하여 스니핏을 추출하는 방법에 대하여 설명하다. 4장에서는 실험 및 분석결과를 보이고, 5장에서 결론을 맺는다.

Ⅱ. 관련 연구

2.1. 스니핏 추출

다음은 현재 많이 연구되고 있는 스니핏에 대한 관련연구이다. Ko등[2]은 웹 스니핏 생성을 위하여 의사연관 피드백과 통계적 질의 확장을 이용하였다. 이들의 방법은 통계 관련 가중치를 이용한 질의확장과 의사연관 피드백을 이용한 질의기반 요약방법을 이용하여서 사용자의 피드백 없이 스니핏을 생성한다. 그러나 초기 질의가 편향된 정보를 포함하고 있을 시 사용자의 의도와는 관계가 적은 스니핏을 추출할 수 있다. Li와 Chen [3]은 통계 언어 모델을 이용한 개인화 문자 스니핏 추출 방법을 제안하였다.

이들의 방법은 은닉 마코프 모델과 확률 기반의 단어의 전후관계 분석방법을 이용하여서 사용자의 의도를 가장 잘 만족 시키는 스니핏을 추출한다. Penin등 [4]은 시맨틱 웹 검색엔진을 위한 스니핏 생성 방법을 제안하였다. 이들의 방법은 이전 온톨로지 요약 방법을 확장하여 온톨로지 스니핏을 지원할 수 있도록 했다. 그러나 이들의 방법은 지원 서술구조 (RDF, resource description framework)기반의 유사도 척도를 사용하기 때문에 원본 자료의 구조에 제약을 받는다. Huang등 [5]은 XML 검색 결과를 위한 스니핏 추출 시스템을 제안하였다. 이들 방법 역시 원본자료의 구조에 많은 영향을 받는다. Turpin등 [6]은 높은 질의 스니핏을 빠르게 생성할 수 있는 새로운 알고리즘 및 단일 파일 구조의 압축 방법을 제안하였다. 이들의 방법은 웹 검색엔진을 위하여 질의 기반의 스니핏을 효율적으로 빠르게 생성할 수 있다.

본 논문의 저자들은 이전에 의사연관 피드백과 퍼지 연관을 이용한 개인화된 스니핏 추출방법을 제안하였다. 이전에 제안된 방법은 질의에 포함된 용어와 페이지에 포함된 용어간의 퍼지관계를 이용함으로 사용자의 질의와 확장된 질의들 간의 표현에는 제약사항을 가지고 있다[7].

2.2. 연관피드백

연관 피드백의 기본이 되는 방법은 Rocchio의 방법으로, 원래의 질의 벡터 $\vec{q}$에 연관된 문장에 대응하는 벡터의 가중치 합을 단순히 더하고, 비연관 문장의 가중치 합을 빼는 방법으로 식(1)과 같다[8, 9].

$$\overrightarrow{q^{new}} = \alpha\vec{q} + \beta \sum_{\forall \overrightarrow{d_{*j}} \in D_+} \overrightarrow{d_{*j}} - \gamma \sum_{\forall \overrightarrow{d_{*j}} \in D_-} \overrightarrow{d_{*j}} \qquad (1)$$

여기서, $\overrightarrow{q^{new}}$는 새롭게 확장된 질의이고, α, β, γ는 조정이 가능한 매개변수들로 일반적으로 $\alpha=\beta=\gamma=1$로 고정하여 사용하며, $\overrightarrow{d_{*j}}$는 j번째 문장의 벡터이다. D+와 D-는 질의에 대한 각각 연관 문장 및 비연관 문장 집합으로서, 사용자에 의서 수동으로 선택되면 연관 피드백이라 하고, 사용자의 개입 없이 자동으로 선택되면 의사연관 피드백이라 한다.

2.3. 퍼지 함의 연산자

퍼지 함의 연산자 (fuzzy implication operator) 는 크리스프 함의 연산자 (crisp implication operator) 를 확장하여 퍼지에 적용한 것으로서, 크리스프 함의 연산자는 {0,1}×{0,1}→{0,1} 로 정의되는데 반해, 퍼지 함의 연산자는 [0,1]×[0,1]→[0,1] 로서 단위 구간의 다치 논리로 확장된 것이다. 퍼지 함의 연산자의 종류는 무수히 많으며 대표적인 Kleene-Diense 퍼지함의 연산자의 예는 다음 식(2)와 같다[10, 11, 12].

$$a \rightarrow b = (1-a) \vee b = \max(1-a, b),$$
$$a = 0 \sim 1,\ b = 0 \sim 1 \qquad (2)$$

집합이론에서 "$A \subseteq B$"는 "$\forall x, x \in A \rightarrow xB$"와 같고 "$A \in \wp(B)$"와도 같다. 여기서 $\wp(B)$ 는 B 의 멱집합 (power set) 이다. 따라서 퍼지 집합에서의 "$A \in B$ 인 정도" 는 $A \in \wp(B)$ 인 정도이므로 $\mu_{\wp(B)}A$ 로서 나타낼 수 있으며 다음과 같이 정의된다.

(정의1) 퍼지 함의 연산자 → 와 크리스프 전체집합 U 의 퍼지 집합 B 가 주어진 상태에서 B 의 퍼지 멱집합의 멤버십 함수 $\mu_{\wp B}$ 는 다음과 같이 주어진다.

$$\mu_{\wp B}A = \bigwedge_{x \in U} (\mu_A x \rightarrow \mu_B x) \ \blacklozenge$$

(정의2) U_1, U_2, U_3는 유한한 전체 집합이라 하고, R 은 U_1 에서 U_2로의 퍼지관계이고, S 는 U_2에서 U_3로의 퍼지관계이다. 즉, R 은 U_1U_2 의 퍼지 부분집합이고 S 는 U_2U_3 의 퍼지 부분집합이다. 퍼지 관계 곱은 a U_1 이고 c U_3 일 때, a 가 c 에 관련되어 있는 정도를 나타낼 사용되는 퍼지연산이다. U_1 에서 U_2 로의 퍼지관계인 삼각논리곱 ,◁, 는 다음과 같이 정의된다.

$$(R \triangle S)_{ik} = \frac{1}{N_j} \sum_j (R_{ij} \rightarrow S_{jk})$$

이것을 퍼지 관계곱 (fuzzy relational products) 이라 한다. ◆

(정의3) 퍼지 함의 연산자는 주어진 문제의 범주에 따라 달라진다. $a \in U_1$ 에 대한 후위집합 (afterset) aR 는 a와 연관된 $y \in U_2$로 구성된 U_2의 퍼지 부분집합이며 그 멤버십 함수는 $\mu_{aR}(y) = \mu_R(a,y)$ 로 주어진다. $c \in U_3$ 에 대한

전위집합 (foreset) Sc 는 c 에 연관된 $y \in U_2$ 로 구성된 U_2 의 퍼지 부분집합이며 그 멤버십 함수는 $_{Sc}(y) = _S(y,c)$ 로 주어진다. aR 이 Sc 의 부분집합인 평균정도는 $y \in aR$ 의 멤버십 정도가 $y \in Sc$ 의 멤버십 정도를 함의하는 평균 정도로서 다음과 같이 정의된다.

$$\pi_m(aR \subseteq Sc) = \frac{1}{N_{U_2}} \sum_{y \in U_2} (\mu_{aR}(y) \rightarrow \mu_{Sc}(y)) \qquad (3)$$

여기서 π_m 은 평균 정도를 나타내는 함수이다. 위의 평균 정도는 $R \triangleleft S$ 에 의해서 a 가 c 에 관련되는 정도를 나타낼 수 있다[10, 11, 12].

Ⅲ. 제안방법

본 논문에서 제안한 스니핏 추출 과정은 다음 그림1과 같이 전처리, 연관 피드백, 스니핏 추출로 구성된다. 전처리단계에서는 검색 문서를 전처리하여서 용어-문장 빈도행렬을 구성한다. 연관 피드백 단계에서는 사용자의 초기질의와 용어-문장 빈도행렬을 이용하여서 질의를 확장하여 집합을 구성한다. 스니핏 추출단계에서는 확장된 질의 집합과 용어-문장 빈도행렬에 퍼지함의 연산자를 이용하여 질의들 간의 포함관계를 계산하고, 질의 간의 포함관계로 부터 문장의 포함관계를 유도하여 스니핏을 추출한다.

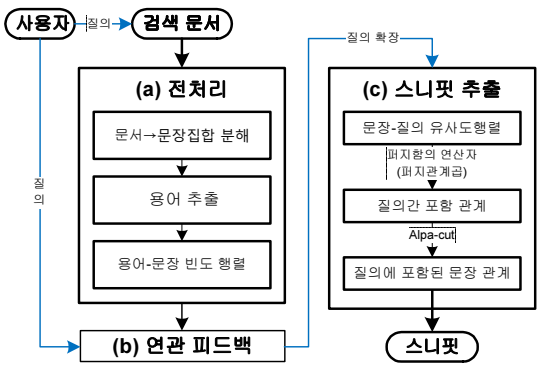


그림 1. 제안된 스니핏 추출 방법
Fig. 1 Proposed method for Snippet extraction

3.1. 전처리

본 논문에서는 전처리는 영문문서를 처리하는 방법[8, 9]을 기준으로 설명하며, 한글 문서는 전처리 단계 중 용어추출에서만 한글 형태소분석 도구[13]를 사용하여 용어를 추출하여 용어-문장 빈도행렬을 구성한다. 그림 1(a)의 전처리 단계로 주어진 문서집합을 문장집합으로 분류하고, 불용어 제거와 어근을 추출한 다음 용어집합을 생성하며, 마지막으로 용어-문장 빈도 행렬을 생성한다[4, 5, 14]. 불용어 제거는 Rijsbergen의 불용어 목록[14]을 이용하여서 목록에서 정의하고 있는 무의미한 용어들을 제거하고, 어근추출은 Porter의 어근추출 알고리즘[14]을 이용하여서 중심이 되는 용어인 어근으로 변환하여 용어집합을 구성한다. 용어집합을 이용하여 용어-문장 빈도 행렬을 생성한다.

생성된 용어-문장 빈도 행렬 d_i 는 $[\, t_{1i}, t_{2i}, \cdots, t_{ni} \,]^T$로 i 번째 문장의 용어빈도이다. 여기서 요소 t_{ij}는 j번째 문장에서 출현한 i번째 용어의 빈도이고, T는 전치행렬을 나타낸다[8, 14].

3.2. 연관 피드백

그림1(b)의 연관 피드백 단계는 사용자의 질의가 스니핏에 충분히 반영할 수 있도록 질의를 확장하여 질의 집합을 구성한다. 본 논문에서 사용하는 연관 피드백 방법은 의사연관 피드백 방법으로, 질의와 문장사이의 유사도를 계산하고, 기본 질의와 유사도가 가장 높은 상위 10개의 문장들 각각에 대하여 질의를 확장하고 확장된 질의 집합을 구성한다. 본 논문에서는 확장되는 질의의 개수를 10개로 한정하였다. 이는 질의 확장에 사용되는 문장의 개수가 너무 많으면 사용자가 원하는 주제를 너무 포괄적으로 포함함으로써 스니핏의 요약내용이 모호한 의미를 나타낼 수 있으며, 문장의 개수가 너무 적으면 사용자의 의도에 너무 협소한 스니핏 요약 결과를 나타낼 수 있기 때문이다.

질의와 문장 간의 유사도 계산은 식(4)의 코사인 유사도를 이용한다[8, 14].

$$sim(d_{*j}, q) = \frac{\sum_{i=1}^{n} d_{ij} \times q_i}{\sqrt{\sum_{i=1}^{n} d_{ij}} \times \sqrt{\sum_{i=1}^{n} q_i}} \qquad (4)$$

여기서 d_{*j}는 j번째 문장의 벡터를 나타내고, q는 질의 벡터를 나타내며, n은 용어의 수를 나타낸다.

의사연관 피드백은 사용자의 연관문서 판단이 없기 때문에 연관 피드백과는 달리 비 연관 문서를 판단 할 수 없다. 이 때문에 본 논문에서는 일반적인 의사연관 피드백에 많이 사용하는 식(5)와 같은 양의 연과 피드백을 사용한다.

$$\overrightarrow{q^{new}} = \vec{q} + \sum_{\forall \overrightarrow{s_{*j}} \in D_+} \overrightarrow{s_{*j}} \quad (5)$$

여기서, $\overrightarrow{q^{new}}$는 의사연관 피드백을 이용하여 새롭게 확장된 질의 벡터이고, $\vec{q}$는 사용자 질의 벡터이다. s는 식(4)를 이용하여서 추출된 질의와 유사도가 가장 높은 문장 벡터이다.

3.3. 스니핏 추출

그림1(c)의 퍼지 함의 연산자를 이용한 스니핏 추출 단계는 다음과 같다. 식(4)의 코사인 유사도를 이용하여서 그림1(b)의 연관 피드백단계에서 확장된 질의 집합과 문장집합 사이의 유사도를 계산하여 질의-문장 유사도 행렬을 구성한다. 구성된 질의-문장 유사도 행렬에 식(2)의 Kleene-Diense 퍼지함의 연산자를 기반으로 한 식(3)의 퍼지 관계곱을 계산하여 질의와 질의 간의 포함관계를 계산한다. 퍼지 이론의 α-cut을 이용하여 질의간의 포함관계로 조절하고, 질의에 포함된 문장 간의 관계를 이용 스니핏을 추출한다. α-cut 은 퍼지 값을 α를 기준으로 0이나 1로 변환하는 식으로, 퍼지 값이 α보다 크면 1로, 작으면 0으로 변환한다. 다음은 퍼지함의 연산자를 이용하여 스니핏을 추출하는 예이다.

예) 표1과 그림2는 질의와 문장 간의 퍼지 함의 연산을 이용하여 스니핏을 추출하는 예이다. 표1(a)는 연관 피드백에 의해서 확장된 질의 집합과 문장 집합 간의 코사인 유사도 행렬 R로, q는 질의를 s는 문장을 나타낸다. 표1(b)는 표1(a)와 식(3)을 이용하여 질의와 질의 간의 퍼지 관계곱을 나타내며, R은 문장-질의 행렬이며, S는 R의 전치행렬(transpose)이다. 표1(c)와 표1(d)는 각각 표1(b)를 $\alpha = 0.9$와 $\alpha = 0.8$의 값으로 α-cut한 결과이다.

표 1. 질의와 문장 간의 퍼지 함의 연산
(a) R (b) R ◁ S (c) α >= 0.9 (d) α >= 0.8
Table. 1 fuzzy implication operator between query and sentence
(a) R (b) R ◁ S (c) α >= 0.9 (d) α >= 0.8

	q_1	q_2	q_3	q_4
s_1	0.9	0.1	0.4	0
s_2	0.1	0.2	0.31	0
s_3	0	0	0.2	0
s_4	0.1	0	0.1	0.1
s_5	0.01	0	0	0.2

(a)

	q_1	q_2	q_3	q_4
q_1	0.9225	0.5	0.7333	0.9450
q_2	0.85	0.85	0.85	0
q_3	0.83	0.64	0.7475	0.9
q_4	0.85	0	0.9	0.85

(b)

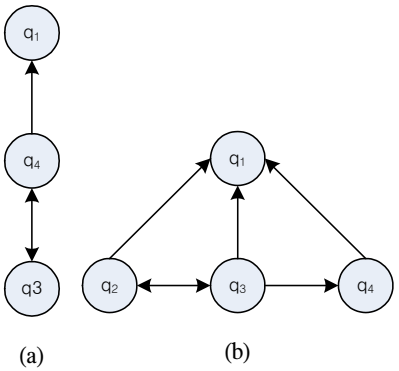
	q_1	q_2	q_3	q_4
q_1	1	0	0	1
q_2	0	0	0	0
q_3	0	0	0	1
q_4	0	0	1	0

(c)

	q_1	q_2	q_3	q_4
q_1	1	0	0	1
q_2	1	1	1	0
q_3	1	0	0	1
q_4	1	0	1	1

(d)

그림2는 본 논문의 제안방법에 대한 이해를 돕도록 표1(c)와 표1(d)를 그래프로 나타낸 것이다. 그림2를 통하여 질의 간의 포함관계를 알 수 있다. 여기서 α-cut의 값이 클수록 질의에 포함되는 질의 수가 적어지며, 많을 수 록 포함되는 질의 수가 많아지는 것을 알 수 있다.



(a) (b)

그림 2. 표1(c)와 (d)의 그래프 표현
(a) 표1(c)의 그래프 (b) 표1(d)의 그래프
Fig. 2 Representation of graph of Table 1 (c) and (d)
(a) graph of Table 1(c) (b) graph of Table 1(d)

그림2(a)로 부터 스니핏 추출 예는 다음과 같이 설명할 수 있다. 표1(a)와 그림2(a)를 통하여 질의에 포함되는 문장과 유사도를 알 수 있다. 그림2(a)에서 질의에 포함되는 문장은 $q_1 = \{s_1, s_2, s_4, s_5\}$, $q_4 = \{s_1, s_3\}$, $q_3 = \{s_1, s_2, s_3, s_4\}$이다. 즉, 그림2(a)로부터 스니핏으로 추출되는 문장은 $\{s_1, s_3, s_2, s_4, s_5\}$와 같이 코사인 유사도 순으로 정렬한다. 스니핏의 추출 크기가 정해진 경우에는 유사도가 높은 순으로 추출한다.

Ⅳ. 실험 및 평가

본 논문에서는 실험 자료로 야후코리아 검색엔진[15]에서 20건의 질의에 대하여 각각 100건의 기사를 검색하여 실험 자료로 사용하였다. 제안 방법을 평가하기 위하여 세 명의 평가자가 문서를 수동으로 50개 이내의 단어로 요약하였다. 즉, 수동으로 요약한 요약문과 제안방법으로 요약된 요약문들 간의 정확률, 재현율, F-measure를 비교 평가 하였다. 성능 평가 척도는 문서요약에서 주로 사용되는 정확률(Precision), 재현율(Recall), F-measure를 척도로 이용하였다[8, 9, 14]. 평가 척도는 다음 식(6)과 같다.

$$R = \frac{|S_{man} \cap S_{sum}|}{|S_{man}|}, \qquad P = \frac{|S_{man} \cap S_{sum}|}{|S_{sum}|}, \qquad F = \frac{2RP}{R+P} \tag{6}$$

여기서 S_{man}, S_{sum}은 각각 사람과 제안된 방법에 의하여 선택된 문장이다.

제안방법의 성능평가는 그림3와 같이 여섯 가지 방법에 대한 평가척도를 비교하였다. 그림3에서 TFIDF [8, 14]는 정보검색 및 문서요약에서 많이 사용하는 코사인 유사도를 이용한 방법이며. PRF는 의사연관 피드백을 사용한 방법[8, 9], TPRF는 의사연관 피드백의 초기 질의를 문서의 제목을 이용한 방법[2], KPRF는 Ko[2]가 제안한 방법으로 후보 문장의 중요도 점수에 기반을 둔 의사연관 피드백을 사용한 방법, FA는 의산연관피드백과 퍼지연관을 이용한 방법으로 이전에 저자들에 제안한 방법이며[7], FIO는 본 논문에서 제안한 방법으로 연관 피드백과 퍼지 함의 연산자를 이용한 방법이다.

비교평가결과 그림3에서 보는 것과 같이 제안 방법인 FIO의 평균 재현율, 정확률, F-measure가 TFIDF에 비하여 12%, 33%, 26.04%가, PRF에 비해서는 10%, 15%, 12.91%가, TPRF에 비해서는 6%, 11%, 8.88%가, KPRF에 비해서는 5%, 5%, 5.01%가, FA에 비해서는 3%, 3%, 3.01%가 더 높다.

성능 평가 결과 제안방법인 FIO가 가장 좋은 결과를 나타내며, 다음으로 FA, KPRF, TPRF, PRF, TFIDF 순으로 평가 되었다. 이는 단순히 문장 간의 유사도를 이용하여서 스니핏을 추출하는 TFIDF방법 보다는 문장 간의 유사도를 이용하여 질의를 확장하여서 스니핏을 추출하는 PRF방법이 더 좋은 성능을 나타내는 것을 알 수 있다. 또한 일반적인 PRF방법보다 의사연관 피드백에 문서의 주제를 초기 질의로 사용한 TPRF방법이나 문장의 중요도를 질의에 반영한 KPRF 방법이 더 좋은 성능을 나타낸다.

특히 제안 방법인 FIO방법은 문장 간의 포함관계를 반영하는 FA에 비해서 질의와 확장된 질의들 간의 포함관계를 반영하는 스니핏을 추출함으로써 가장 좋은 성능을 보이는 것으로 분석된다.

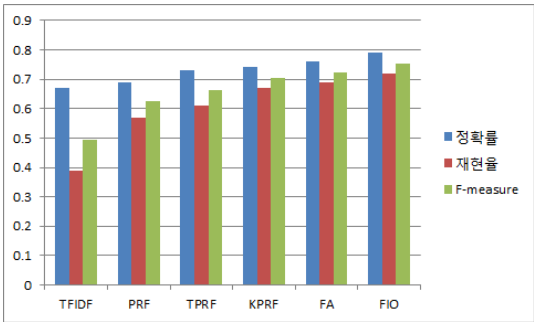


그림 3. 평가방법에 대한 성능비교 결과
Fig. 3 Evaluations results of performance comparison

Ⅴ. 결론

본 논문에서는 연관 피드백과 퍼지 함의 연산자를 이용한 새로운 스니핏 추출 방법을 제안하였다. 제안방법은 연관 피드백에 의해서 확장된 질의집합에 퍼지 함의 연산자를 적용하여 질의 간의 포함관계가 반영됨으로써 사용자의 의도가 의미적으로 더 잘 포함되는 스니핏을 추출할 수 있도록 하였다. 실험 결과 제안방법이 정보

검색에 많이 사용하는 문장의 유사도에 의한 방법이나 연관 피드백을 기반으로 한 방법, 일반적인 퍼지를 이용한 방법보다 더 좋은 성능을 보였다.

참고문헌

[1] G. Manolache, “Index-based Snippet Generation”, Master’s Thesis, 2008.

[2] Y. J. Ko, H. K. An, J. Y. Seo, ″Pseudo-relevance feedback and statistical query expansion for web snippet generation,″ Information Processing Letter. Vol. 109, pp.18-22, 2008.

[3] Q. Li, Y. P. Chen, ″Personalized text snippet extraction using statistical language models,″ Pattern Recognition, Vol. 43, pp.378-386, 2010.

[4] T. Penin, H. Wang, T. Tran, Y. Yu, ″Snippet Generation for Semantic Web Search Engine,″ In proceeding of ASWC, LNCS 5367, pp.493-507, 2008.

[5] Y. Huang, Z. Liu, ″Query Baised Snippet Generation in XML Search,″ In proceeding of SIGMOD, pp.??, 2008.

[6] A. Turpin, Y. Tsegay, D. Hawking, H. E. Williams, ″Fast Generation of Result Snippets in Web Search,″ In proceeding of SIGIR, pp.127-134, 2007.

[7] 박선, 조광문, 양후열, 이성로, “의사연관 피드백과 퍼지연관을 이용한 개안화 문서 스니핏 추출 방법”, 전자공학회 논문지, 제49권 SP편 제2호, pp.137-142, 2012.

[8] B. Y. Ricardo, R. N. Berthier, ″Moden Information Retrieval,″ ACM Press, 1999.

[9] S. Chakrabarti, ″mining the web: Discovering Knowledge from Hypertext Data,″ Morgan Kaufmann Publishers, 2003.

[10] K. W. Oh and W. Bandler, ″Properties of fuzzy implication operatoers”, International Journal of Approximate Reasoning, Vol. 1, No. 3, pp.273-285, 1987.

[11] Bandler, W. and Kohout, L., “Fuzzy Power Sets and Fuzzy Implication Operations,” Fuzzy Set and Systems, Vol.4, No.1, pp. 13-30, 1980.

[12] Bandler, W. and Kohout, L., “Semantics of Implication Operators and Fuzzy Relational Products,” International Journal of Man-Machine Studies, Vol. 12, pp.89-116, 1980.

[13] 한경남, 남경완, “한국어정보처리 입문 : 컴퓨터가 우리말을 이해하려면?”, 커뮤니케이션북스, 2007.

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

[15] 야후 코리아, www.yhaoo.co.kr, 2011.

저자소개

박선(Park Sun)

1996년 전주대학교(학사)
2001년 한남대학교 정보산업대학원(석사)
2007년 인하대학교 대학원(박사)
2008～2009년 호남대학교 컴퓨터공학과 전임강사
2010년 전북대학교 전기전자정보 인력양성사업단 박사후과정
2011년～현재 목포대학교 정보산업연구소 전임연구교수

※관심분야 : 정보검색, 데이터마이닝, 데이터베이스, 해양생물 IT정보융합

심천식(Shim Chun Sik)

1995년 인하대학교(학사)
1997년 인하대학교(석사)
2003년 인하대학교(박사)
2008년 9월～현재 목포대학교 조선공학과 교수

※관심분야 : 조선IT정보융복합

이성로(Lee Seong Ro)

1987년 고려대학교(학사)
1990년 한국과학기술원(석사)
1996년 한국과학기술원(박사)
1997년 9월~현재 목포대학교
정보전자공학과 교수

※관심분야 : 디지털통신시스템, 이동 및 위성통신 시스템, USN/텔레미틱스응용분야, 임베디드시스템