퍼지 추론을 이용한 소수 문서의 대표 키워드 추출

Representative Keyword Extraction from Few Documents through Fuzzy Inference

노순억*, 김병만*, 허남철**
*금오공과대학교 대학원 컴퓨터공학과
**대구미래대학 컴퓨터정보처리학과

Sun Ok Rho*, Byeong Man Kim*, Nam Chul Huh**
*Dept. of Computer Engineering, Kumoh National University of Technology.
**Dept. of Computer Information Processing, Daegu Mirae College.
(sorho@cespc1.kumoh.ac.kr)

ABSTRACT

In this work, we propose a new method of extracting and weighting representative keywords(RKs) from a few documents that might interest a user. In order to extract RKs, we first extract candidate terms and then choose a number of terms called initial representative keywords (IRKs) from them through fuzzy inference. Then, by expanding and reweighting IRKs using term co-occurrence similarity, the final RKs are obtained. Performance of our approach is heavily influenced by effectiveness of selection method of IRKs so that we choose fuzzy inference because it is more effective in handling the uncertainty inherent in selecting representative keywords of documents. The problem addressed in this paper can be viewed as the one of calculating center of document vectors. So, to show the usefulness of our approach, we compare with two famous methods - Rocchio and Widrow-Hoff - on a number of documents collections. The results show that our approach outperforms the other approaches.

Keywords : keyword extraction, fuzzy inference, user preference

Ⅰ. 서 론

웹 검색 엔진 혹은 다양한 정보 검색 시스템을 이용하는 일반 사용자는 자신이 원하는 내용에 가장 적합한 정보를 찾고자 관심 대상 영역에 대한 제한된 어휘력과 전문성을 바탕으로 검색 질의어를 구성한다. 마찬가지로 정보 필터링 시스템 이용시 사용자는 적절한 정보를 추천 혹은 제공받고자 자신의 프로파일에 관심사항을 기술한다.

검색 시스템의 경우 제공된 질의어로 검색 기능을 수행하고 그 결과에 대해서 사용자로부터 피드백을 받거나 검색된 결과를 이용해 자동으로 질의어를 수정하고 중요도를 재산정하는 등의 부가 기능들을 수행함으로써 사용자에게 편의성을 제공하고 검색 효율을 높이고 있다[1, 3]. 정보 필터링 시스템 역시 위와 비슷한 성격의 프로파일 수정 과정들을 가진다.

사용자에 의한 적절한 프로파일(질의어) 작성은 사용자에게 부담을 줄 수 있고 용어 불일치 문제로 인한 부적절한 필터링(검색) 결과를 가

져 올 수 있다. 따라서 사용자의 프로파일 작성의 부담과 용어 불일치 문제의 수위을 낮추기 위해서 사용자로부터 관심 내용과 유사한 문서 집합을 제공 받아 이를 활용하는 것도 하나의 해결 방법이 될 수 있다. 이 경우에 발생되는 문제점은 제공된 문서 집합으로부터 사용자를 대신해서 대표 용어를 추출하고 이들에게 어느 정도의 중요도를 부여 할 것인가이다. 본 논문에서는 위의 문제 해결을 위한 새로운 방법을 제시한다.

II. 본 론

예제(학습) 문서들로부터 대표 용어 확장 및 가중치 재산정에 필요한 사용자의 관심 내용을 가장 잘 대변하는 초기 대표 용어의 선택이 무엇보다 중요하다. 이들 초기 대표 용어들과 각각의 예제 문서 내에 존재하는 후보 용어들과의 발생 빈도 유사도 계산이 서로 이루어짐으로 선택 방법과 기준이 성능에 큰 영향을 미친다. 특정 용어의 중요도 계산에 사용되는 입력 정보(예: TF, DF, IDF)들은 정량적으로 정확히 해석될 수 없는 부정확하고 불확실한 특성을 내포하고 있다. 따라서 본 논문에서는 이러한 불확실성의 문제 해결에 효과적인 퍼지 추론을 적용하여 후보 용어들의 가중치을 계산하고 이 값들에 따라 선택 우선 순위를 부여하였다(2.1참조).

초기 대표 용어들은 선택 우선 순위에 따라서 각각의 예제 문서가 적어도 하나의 대표 용어를 포함할 때까지 확장되어 진다(2.2참조). 위의 결과로 초기 대표 용어들이 생성되면 다음 단계로 각각의 예제 문서 내의 후보 용어들과의 발생 빈도 유사도를 이용한 후보 용어들의 가중치 재산정이 수행된다(2.3참조). 기존의 학습 문서 집합을 대표하는 용어들의 가중치 부여 방법들(Rocchio, Widrow-Hoff)에서는 구성 용어들간의 어떠한 관련성을 계산에 반영하고 있지 않다[2]. 이에 초기 대표 용어들이 사용자의 관심 내용을 가장 잘 나타내고 있다는 가정 아래 이들 용어들과 후보 용어들과의 발생 빈도의 유사성을 하나의 관련성으로 두고 중요도 계산에 이를 반영함으로써 분류 성능의 향상을 도모하였다[1].

대표 용어들의 자동 확장 단계에서는 기존 비교 방법들과 동일하게 모든 후보 용어들을 확장에 포함시켰으며 용어 가중치 재산정의 결과로 수정된 각각의 예제 문서들의 가중치 벡터들과 정보 검색 분야의 질의 용어 가중치 계산식[1, 3]을 사용한 초기 대표 용어들의 가중치 벡터를 전부 합산하여 최종적으로 예제 문서 집합의 대표 가중치 벡터를 구성하였다.

2.1 퍼지 추론을 이용한 대표용어 중요도계산

▣ TF(Term Frequency)

각 용어의 발생 빈도수는 퍼지 계산에 사용되어지기 위해 정규화(NTF)되어야 하며 아래의 (식 1)을 사용하였다.

$$NTF_i = \frac{\frac{TF_i}{DF_i}}{\underset{j}{Max}\left[\frac{TF_j}{DF_j}\right]} \quad (1)$$

TF_i : 예제 문서 집합에서 i 번째 단어의 발생 빈도수

DF_i : 예제 문서 집합에서 i 번째 단어를 포함하는 문서의 수

▣ DF(Document Frequency)

각 용어의 예제 문서 집합 내에서의 문서 발생 빈도수를 나타내며 TF와 마찬가지로 아래의 (식 2)을 사용하여 정규화(NDF) 하였다

$$NDF_i = \frac{\frac{DF_i}{TD}}{\underset{j}{Max}\left[\frac{DF_j}{TD}\right]} \quad (2)$$

TD : 예제 문서의 수

DF_i : 예제 문서 집합에서 i 번째 단어를 포함하는 문서의 수

▣ IDF(Inverse Document Frequency)

각 용어의 전체 예제 문서 집합 내에서의 역문헌 빈도수를 나타내며 아래의 식 (3)을 사용하여 정규화(NIDF) 하였다.

$$NIDF_i = \frac{IDF_i}{\underset{j}{Max}\left[IDF_j\right]} \quad (3)$$

IDF_i : i 번째 단어의 역문헌 빈도수

▣ 용어별 중요도 계산

그림 1은 퍼지 추론을 위하여 사용된 입출력 변수들을 나타내고 있다. 용어별로 구해진 NTF, NDF, NIDF 값들을 퍼지 추론에 적합한 형태로 퍼지화 시켜야 한다.

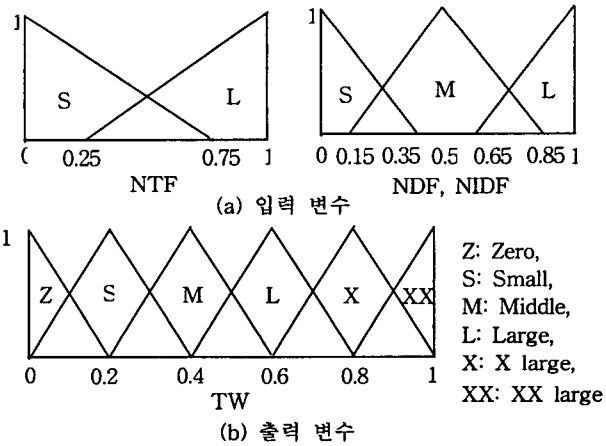

그림 1 : 퍼지 입출력 변수

표 1은 NTF 퍼지 입력값의 소속 정도에 따라 두 부분으로 나누어 규칙들을 표현하고 있다. NTF, NDF, NIDF 퍼지 입력값을 위의 결과로 생성된 18개의 추론 규칙별로 이들의 전건부의 소속 함수에 적용시킨다. 각각의 소속 정도가 구해지면 이들 중에서 최소(min)값을 취한다. 그 결과 규칙별로 하나씩의 퍼지 값이

생성되며 이 퍼지 값들을 퍼지 출력 변수 TW에 따라 6개의 그룹으로 분류하고 그룹별로 해당 그룹에 속한 퍼지 값들 중 최대(max)값을 취하여 총 6개의 퍼지 값들을 생성한다. 최종적으로 이들 6개의 퍼지 값들을 무게중심법(center of gravity)으로 비퍼지화(defuzzification)한 값이 해당 용어의 중요도 값으로 결정되어진다.

표 1 : 퍼지 추론 규칙

NIDF / NDF	S	M	L
S	Z	Z	S
M	Z	M	L
L	S	L	X

NTF = S

NIDF / NDF	S	M	L
S	Z	S	M
M	S	L	X
L	S	X	XX

NTF = L

2.2 초기 대표 용어 선택

퍼지 추론을 통해 후보 용어들의 중요도 값들이 계산되어지면 이 값들에 따라 선택 우선순위를 부여한다. 각각의 예제 문서가 적어도 1개의 초기 대표 용어를 포함해야 한다는 제약사항을 두었다. 그림 2는 초기 대표 용어 선택 알고리즘을 보여주고 있다.

```
Input: DS ( Example Documents Set )
       TS ( Candidate Terms Set )
1] Procedure get_ITS(DS, TS)
2] ITS: Initial Representative Terms Set,
     initialized to empty.
3] TS': Temporary Terms Set, initialized to TS.
4] d, t: Document and Term element respectively.
5] Repeat
6]   Select a document element as d from DS.
7]   Repeat
8]     Select the highest element as t in TS'
         according to the weight.
9]     If t appears in d and not member in ITS
         Then Add t to ITS.
10]    Remove t from TS.
11]  Until t appears in d.
12]  Remove d from DS.
13]  Assign TS to TS'.
14] Until DS is empty.
15] Return ITS.
```

그림 2 : 초기 대표 용어 선택 알고리즘

2.3 용어 가중치 계산과 대표용어 자동확장

초기 대표 용어들의 선택이 완료되면 이들 용어들과 각 문서내에서의 후보 용어와의 발생 빈도 유사도를 아래의 식 (4)를 이용하여 계산하게 된다.

$$Rd_{il}(K,\ t_i) = 1 - \log_p\left(\sqrt{\frac{\sum_{j=1}^{n}(kf_{jl} - tf_{il})^2}{n}}\right) \quad (4)$$

$Rd_{il}(K,\ t_i)$: 문서 l 에서 후보 용어 t_i 와 초기 대표 용어간의 관련 정도
kf_{jl} : 문서 l 에서 초기 대표 용어 j의 발생 빈도수
tf_{il} : 문서 l 에서 후보 용어 i의 발생 빈도수
K : 전체 초기 대표 용어들
n : 초기 대표 용어들의 개수
p : 조정 상수

후보 용어와 초기 대표 용어들간의 관련 정도가 산정되면 다음 단계로 이를 반영하면서 문서 집합에 대한 후보 용어의 가중치를 식 (5)를 이용하여 계산하게 된다

$$wt_i = \sum_{l=1}^{n}(wt_{il} \times Rd_{il}) \quad (5)$$

$$wt_{il} = TF_{il} \times IDF_i$$

wt_i : 후보 용어의 문서 집합에서의 가중치
wt_{il} : 문서 l 에서 후보 용어 i의 가중치
Rd_{il} : 문서 l 에서 후보 용어 i와 초기 대표 용어들간의 관련 정도
TF_{il} : 문서 l 에서 후보 용어 i의 발생 빈도수
IDF_i : 후보 용어 i의 역문헌 빈도수
n : 문서 집합내의 문서 개수

2.4 실험 및 결과

본 논문의 제안 방법의 유용성을 평가하기 위해서 기존의 대표적인 선형 분류기인 Rocchio, Widrow-Hoff 알고리즘들과의 문서 분류 성능을 비교해 보았다. 예제 문서 집합으로부터 대표 용어들을 추출하고 이들에게 가중치를 부여하는 문제는 위의 비교 대상 알고리즘들이 학습 문서 집합의 중심(center) 벡터를 구성하는 것과 성격이 같다.

실험 문서 집합으로는 Reuters-21578을 선택하였다. 본 논문에서는 Reuters-21578의 ApteMod 버전을 사용했고 라벨이 없는 문서들은 제외시켰다. 실험 대상으로 소수 예제 문서 집합들을 준비하고자 테스트 문서 집합과 학습 문서 집합에 적어도 하나의 문서를 각각 포함하고 있는 범주(category)들을 선택(총 90개)한 후 이중에서 학습 문서 개수가 10개~30개인 범주 21개를 마지막으로 선별했다.

테스트 문서 집합은 3019개의 문서들을 포함하고 있다. 용어의 역문헌 빈도수(IDF)값을 구하기 위해 90개의 범주들에 속하는 7770개의 학습 문서 집합으로부터 문서 빈도수 정보를 이용하였다. 사용자는 자신의 관심 사항에 부합하는 긍정적 문서 집합(positive documents)만을 제공한다는 가정하에 알고리즘 수행시 부정적 문서(negative documents)들의 정보 이용은 모두 제외시켰다.

비교 대상 알고리즘에 사용된 벡터들의 가중치는 용어의 TF×IDF로 계산하였다. 실험시 사용된 조정 상수(parameter)들의 설정값들은 Rocchio의 경우 $\alpha=0, \beta=1, \gamma=0$로 두었고 Widrow-Hoff의 경우 $\eta=0.25$, $y_i=1$로 두었으며 [2] 본 제안 방법에서 사용되는 식 (4)의 p는 10으로 두었다. 유사도 계산식은 cosine 식을 이용했으며 문서 분류 성능의 측도로서 standard recall, precision를 사용했다

표 2는 11 standard recall levels 에 해당하는 보간(interpolation)절차를 통해 얻어진 정확율들의 평균값들과 비교 알고리즘들에 대한 제안 방법의 성능 향상율들을 각 범주별로 보여주고 있다. 본 제안방법(R.K.E.F)의 성능이 Rocchio 보다 평균 17%, Widrow-Hoff 보다는 평균 14% 향상되었음을 알 수 있다.

표 2 : 21개 범주들에 대한 성능 및 초기 대표 용어 추출 방법 별 성능 -

R.K.E.F(Representative Keyword Extraction throuth Fuzzyinference), R.K.E.R(Rocchio), R.K.E.W(Widrow-Hoff).

	R.K.E.F				
	averaged precision	over Rocchio	over W.H	over R.K.E.R	over R.K.E.W
lumber	0.550	+51.9%	+25.8%	+ 34.4%	+ 49.8%
dmk	0.084	+52.7%	+61.5%	+ 0.0%	+ 0.0%
sunseed	0.451	+18.6%	+18.3%	+ 302.6%	+ 302.6%
lei	0.363	+ 0.0%	+ 0.0%	+ 0.0%	+ 0.0%
soy-meal	0.772	+40.6%	+43.7%	+ 101.0%	+ 101.0%
fuel	0.518	+40.3%	+58.8%	+ 36.3%	+ 51.0%
heat	0.626	+ 9.6%	+10.0%	+ 2.6%	+ 2.6%
soy-oil	0.323	+20.5%	+ 0.6%	+ 62.3%	+ 76.5%
lead	0.614	+13.9%	+10.6%	+ 76.4%	+ 16.2%
strategic-metal	0.120	+15.3%	+11.1%	+ 21.2%	+ 20.2%
hog	0.485	- 5.6%	- 9.1%	- 1.2%	- 11.8%
orange	0.975	+ 5.1%	+ 4.3%	+ 0.0%	+ 0.0%
housing	0.352	- 9.0%	- 5.6%	- 10.2%	- 1.6%
tin	0.986	+ 2.0%	+ 1.7%	+ 0.0%	+ 0.0%
rapeseed	0.575	+24.1%	+28.3%	- 0.8%	- 0.8%
wpi	0.728	- 8.1%	- 9.4%	- 4.9%	- 4.9%
pet-chem	0.308	+ 5.4%	-16.9%	+ 4.0%	-11.7%
silver	0.770	+61.4%	+50.0%	+ 2.2%	+ 2.2%
zinc	0.921	+16.2%	+ 8.2%	+ 0.0%	+ 0.0%
retail	0.194	+ 2.6%	+ 1.5%	+ 3.7%	+ 102.0%
sorghum	0.591	+16.3%	+18.2%	- 4.0%	+ 94.4%
Average	0.530	+17.8%	+14.8%	+ 29.8%	+ 37.5%

표 3 : lumber 범주에서 추출된 적용방법별 초기 대표 용어들

Methods	Initial Representative Keywords
Fuzzy	lumber, softwood, wood, agre
Rocchio	lumber, softwood, timber, guarante, forest
W.H	lumber, timber, guarante, champion, wood

본 제안 방법에서는 초기 대표 용어들의 추출 방법으로 퍼지 추론을 이용하였다. 퍼지 추론의 유용성을 확인하고자 비교 알고리즘들을 초기 대표 용어 추출에 이용하여 각각의 성능을 비교해 보았다. 표 3은 각 알고리즘별로 lumber 문서 집합에서 추출되는 초기 대표 용어들의 예를 보여주고 있다. 실험결과, 퍼지 추론을 이용한 방법에 비해 두 비교 알고리즘을 각각 응용한 방법들은 뚜렷한 성능 향상을 보여주지 않았다. 즉 적절한 핵심용어를 선별하지 못했음을 알 수 있었다. 표 2의 오른쪽 결과를 통해 비교 알고리즘들보다 퍼지 추론 방법이 보다 나은 성능을 보여주고 있음을 알 수 있다.

III. 결 론

본 논문에서는 사용자가 제시한 소수의 문서 집합으로부터 관심 내용을 대표하는 중요 용어들을 추출하고 그들의 가중치를 부여하는 문제에 관하여 퍼지 추론과 용어 발생 빈도수의 유사성을 이용한 가중치 재산정 방법을 적용했다. 방법의 유용성을 보이고자 학습 문서 집합의 중심 벡터를 구하는 대표적인 선형 분류기의 알고리즘들과 성능을 비교했다. 소수의 긍정적 학습 문서 집합들에 대해서 실험한 결과 비교적 우수한 성능향상을 보여줌으로써 본 제안 방법의 유용성을 확인할 수 있었다.

사용자의 관심 내용을 가장 잘 대변하는 초기 대표 용어(핵심 용어)들의 추출 방법은 용어 발생 빈도수의 유사성에 따른 가중치의 재산정에 직접적인 영향을 주고 있다. 사용자의 관심 내용과 거리가 먼 핵심 용어를 추출했을 경우 그 용어를 중심으로 빈도수의 유사성을 보이면서 관련 전문 용어 혹은 유의어 등의 성격을 함께 가지는 후보용어의 가중치가 높게 산정됨으로 그 결과로 생성되는 중심 벡터는 문서 집합의 대표성을 상실하게 된다. 따라서 향상된 실험결과를 통해 핵심 용어들이 올바르게 추출 되었고 사용된 퍼지 추론 방법이 효과적이었음을 확인할 수 있다. 본 제안 방법에 이용된 용어들 간의 발생 빈도수의 유사성 이외에 또다른 여러 가지 관련 정보를 이용할 수 가 있고 이와 관련된 문제 해결를 위해 퍼지 추론 방식이 중요한 기준 혹은 기반 구축 방법으로써 효과적으로 사용 될 수 있을 것이다.

감사의 글 : 본 연구는 한국과학재단 목적기초연구 (2000-1-51200-008-2)지원으로 수행되었습니다.

IV. 참고문헌

[1] Byong Man Kim, Ju Youn Kim, JongWan Kim, "Query Term Expansion and Reweighting using Term Co-Occurrence Similarity and Fuzzy Inference" IFSA/NAFIPS, 2001.

[2] David D. Lewis and Robert E. Schapire and James P. Callan and Ron Papka, "Training algorithms for linear text classifiler", Proceedings of SIGIR-96, 19th ACM International Conference on Research and Development in Information Retrieval, 1996.

[3] R.Baeza-Yates, B. Ribeiro-Neto, Modern Information Retrieval, ACM Press, NY, USA, 1999.