개선된 BK-퍼지정보검색모델(A-FIRM)과 BK-퍼지정보검색모델(BK-FIRM)의 성능 평가

Performance Evaluation of A-FIRM and BK-FIRM

김창민, 김용기
Kim, Chang-Min and Kim, Yong-Gi
경상대학교 컴퓨터과학과 및 전산개발연구소
Dept. of Computer Science, Institute of Computer Research and Development
Kyungsang National University

요 약

퍼지관계 개념을 응용한 BK-퍼지정보검색기법은 형태론에 입각하는 기존의 정보검색기법과는 달리 문서와 용어의 상대적 의미에 근거하는 정보검색 기법이다. 그러나 BK-퍼지정보검색기법은 높은 시간복잡도(time complexity)의 검색 연산을 내재하고 있어 실제 대용량의 정보 검색은 사실상 불가능하다. 본 논문에서는 BK-퍼지정보검색모델(BK-FIRM)의 높은 시간복잡도를 낮추기 위해, 용어집합의 부분집합으로서 그 원소 개수는 상수처럼 작용하는 축소용어집합(reduced term set)을 이용한 개선된 퍼지정보검색모델(A-FIRM)을 제안하고 실제 이를 처리시간과 신뢰도 측면에서 분석 및 비교한다.

1. 소개

과학과 기술 분야의 급속한 발전은 수많은 주제들에 대한 방대한 양의 정보를 생성하는 정보화 사회를 탄생시켰다. 따라서 원하는 정보에 대한 정확하고 빠른 접근은 정보화 사회를 살아가는 현대인들에게 성공의 여부를 결정짓는 중요한 요소가 되었다. 그러나 대용량의 데이터로부터 원하는 정보를 한정된 시간 내에 검색하는 것은 쉬운 일이 아니다. 1960년대 초,

이러한 문제점을 해결하기 위하여 컴퓨터를 이용하여 정보를 검색하는 정보검색(information retrieval) 분야가 확립되었다[1][7].

Bandler와 Kohout의 퍼지정보검색모델은 불리언 정보검색모델을 확장한 것으로서 시소러스 자동 구축 기능, 검색 결과의 퍼지화된 우선순위 제공, 직접 관련 없는 제 3의 개체 유추 검색 등과 같은 장점을 가지고 있다. 그러나 BK-퍼지정보검색모델은 높은 시간복잡도(time complexity)의 검색연산을 내재하고 있어 대용량의 정보 검색이 요구되는 분야 적용은 쉽지 않다.

본 논문에서는 축소용어집합을 이용하여 BK-퍼지정보검색모델의 높은 시간복잡도를 낮추는 개선된 BK-퍼지정보검색모델을 제안하고 이를 처리시간과 신뢰도 측면에서 분석 및 비교한다.

2. BK-퍼지정보검색모델

Bandler와 Kohout의 BK-퍼지정보검색모델은 형태론에 입각한 기존의 정보검색기법과는 달리 문서와 용어의 상대적 의미를 표현하는 퍼지관계와 퍼지관계곱을 이용하는 정보검색기법으로서 자동 시소러소 구축기능과 검색결과의 퍼지화된 우선 순위 제공하는 기능을 기본적으로 가지고 있다. BK-퍼지정보검색모델은 시소러스를 이용하여 주어진 용어의 의미를 확장하는 관계요구(R-request) 연산과 사용자로부터 주어진 질의어를 해석하여 적합한 문서를 검색하는 퍼지검색요구(FS-request) 연산을 제공한다[2][5].

우선 수식(1)(2)과 같이 문서집합 D와 용어집합 T을 정의하고 수식(3)과 같이 문서와 용어와의 퍼지관계 $\tilde{R}$이 존재한다고 가정한다. 이때 임의의 검색식 S가 주어지면 수식(4)과 같이 퍼지관계 $\tilde{R}$을 검색식 S에 대입하여 검색식 S에 대한 적합도를 표현하는 문서의 퍼지집합 $\tilde{D}$를 얻고 수식(5)과 같이 $\tilde{D}$에 $\alpha-cut$을 적용, α-레벨 집합화하여 최종 k'개의 결과 문서를 가지는 집합 D'_α를 구한다[5].

$$D = \{d_1, d_2, \Lambda, d_k\} \tag{1}$$

$$T = \{t_1, t_2, \Lambda, t_n\} \tag{2}$$

$$\tilde{R} = D \times T = \begin{bmatrix} v_{11} & v_{12} & \Lambda & v_{1n} \\ v_{21} & v_{22} & \Lambda & v_{2n} \\ \mathrm{M} & \mathrm{M} & \mathrm{O} & \mathrm{M} \\ v_{k1} & v_{k2} & \Lambda & v_{kn} \\ t_1 & t_2 & \Lambda & t_n \end{bmatrix} \begin{matrix} d_1 \\ d_2 \\ \mathrm{M} \\ d_k \end{matrix} \tag{3}$$

$$\begin{aligned} \tilde{D} &= S(\tilde{R}) \\ &= S\left(\begin{bmatrix} v_{11} & v_{12} & \Lambda & v_{1n} \\ v_{21} & v_{22} & \Lambda & v_{2n} \\ \mathrm{M} & \mathrm{M} & \mathrm{O} & \mathrm{M} \\ v_{k1} & v_{k2} & \Lambda & v_{kn} \\ t_1 & t_2 & \Lambda & t_n \end{bmatrix} \begin{matrix} d_1 \\ d_2 \\ \mathrm{M} \\ d_k \end{matrix} \right) \\ &= \{d_1/s_1, d_2/s_2, \Lambda, d_k/s_k\} \end{aligned} \tag{4}$$

$$D'_\alpha = \{d'_1, d'_2, \Lambda, d'_{k'}\} \tag{5}$$

3. 개선된 BK-퍼지정보검색모델

Bandler와 Kohout가 제안한 BK-퍼지정보검색모델은 확장된 불리언 정보검색모델로서 자동 시소러스 구축, 검색결과의 퍼지화된 우선순위 제공, 직접 관련 없는 제 3의 개체 유추 검색 등과 같은 장점을 제공한다. 그러나 BK-퍼지정보검색모델은 높은 시간복잡도(time complexity)의 검색연산을 내재하고 있어 다양한 분야 적용을 어렵게 한다. 본 논문에서는 개선된 축소용어집합(reduced term set)을 이용하여 BK-퍼지정보검색모델의 시간복잡도를 개선하는 개선된 BK-퍼지정보검색모델을 제안한다.

3.1. 축소용어집합

개선된 BK-퍼지정보검색모델에서는 BK-퍼

지정보검색모델의 시간복잡도를 개선하기 위하여 용어집합의 부분집합으로서 상수개의 원소로 구성된 축소용어집합을 이용한다. n개의 원소로 구성된 임의의 용어집합 T에서 n_r개의 T_r 원소를 추출하여 축소용어집합(reduced term set) 을 생성하는 연산 Ω는 수식(6)과 같이 정의된다. 이때 연산 Ω는 무작위 추출, 규칙에 의한 추출 혹은 인간에 의한 수동 추출 등의 다양한 방법으로 구현 가능하다.

$$T_r = \Omega(T, n_r, \chi, \gamma) \tag{6}$$

3.2. 개선된 BK-퍼지정보검색모델과 검색

개선된 BK-퍼지정보검색모델은 사용자로부터 주어진 질의어를 해석하고 시소러스를 이용하여 확장하여 적합한 문서를 검색하는 검색요구 연산을 제공한다. 우선 수식 (7)(8)(9)과 같이 문서집합 D, 용어집합 T, 문서집합과 용어집합의 퍼지관계 $\tilde{R}$이 주어지고 수식 (10)(11)(12)과 같이 축소용어집합 T_r, 문서집합과 축소용어집합과의 퍼지관계 $\tilde{R}_r$, 시소러스 $\tilde{B}_r$가 정의된다. 사용자로부터 질의어 S가 입력되면 수식 (13)과 같이 질의어를 해석하는 연산 Φ에 의하여 $\tilde{Q}$를 산출하고 수식 (14)과 같이 $\tilde{Q}$와 시소러스 $\tilde{B}_r$을 합성하여 확장된 질의어 $\tilde{Q}_r$을 산출한다. 수식 (15)에서는 퍼지관계곱을 이용하여 $\tilde{R}_r$과 확장된 질의어 $\tilde{Q}_r$을 합성하여 검색결과 $\tilde{O}$를 얻고 수식 (16)에서는 $\tilde{O}$에 $\alpha-cut$을 적용하여 최종 문서 검색결과 $\tilde{O}_\alpha$를 구한다. 이때 연산 o, @는 각각 퍼지관계합성연산[9], 퍼지관계곱연산[2][3][6]을 의미한다.

$$D = \{d_1, d_2, \Lambda, d_m\} \tag{7}$$

$$T = \{t_1, t_2, \Lambda, t_n\} \tag{8}$$

$$\tilde{R} = D \times T = \begin{matrix} \begin{bmatrix} v_{11} & v_{12} & \Lambda & v_{1n} \\ v_{21} & v_{22} & \Lambda & v_{2n} \\ \mathrm{M} & \mathrm{M} & \mathrm{O} & \mathrm{M} \\ v_{k1} & v_{k2} & \Lambda & v_{kn} \end{bmatrix} \begin{matrix} d_1 \\ d_2 \\ \mathrm{M} \\ d_k \end{matrix} \\ t_1 \quad t_2 \quad \Lambda \quad t_n \end{matrix} \tag{9}$$

$$T_r = \Omega(T, n', \chi, \gamma) = \{r_1, r_2, \Lambda, r_{n_r}\} \tag{10}$$

$$\tilde{R}_r = D \times T_r = projection(R, T_r) = \begin{matrix} \begin{bmatrix} v_{11} & v_{12} & \Lambda & v_{1n_r} \\ v_{21} & v_{22} & \Lambda & v_{2n_r} \\ \mathrm{M} & \mathrm{M} & \mathrm{O} & \mathrm{M} \\ v_{k1} & v_{k2} & \Lambda & v_{kn_r} \end{bmatrix} \begin{matrix} d_1 \\ d_2 \\ \mathrm{M} \\ d_k \end{matrix} \\ r_1 \quad r_2 \quad \Lambda \quad r_{n_r} \end{matrix} \tag{11}$$

$$\tilde{B}_r = \tilde{R}^T \times \tilde{R}_r = \begin{matrix} \begin{bmatrix} v_{11} & v_{12} & \Lambda & v_{1k} \\ v_{21} & v_{22} & \Lambda & v_{2k} \\ \mathrm{M} & \mathrm{M} & \mathrm{O} & \mathrm{M} \\ v_{n1} & v_{n2} & \Lambda & v_{nk} \end{bmatrix} \begin{matrix} t_1 \\ t_2 \\ \mathrm{M} \\ t_n \end{matrix} \\ d_1 \quad d_2 \quad \Lambda \quad d_k \end{matrix} @ \begin{matrix} \begin{bmatrix} v_{11} & v_{12} & \Lambda & v_{1n_r} \\ v_{21} & v_{22} & \Lambda & v_{2n_r} \\ \mathrm{M} & \mathrm{M} & \mathrm{O} & \mathrm{M} \\ v_{kn} & v_{k2} & \Lambda & v_{kn_r} \end{bmatrix} \begin{matrix} d_1 \\ d_2 \\ \mathrm{M} \\ d_k \end{matrix} \\ r_1 \quad r_2 \quad \Lambda \quad r_{n_r} \end{matrix}$$

$$= \begin{matrix} \begin{bmatrix} v_{11} & v_{12} & \Lambda & v_{1n_r} \\ v_{21} & v_{22} & \Lambda & v_{2n_r} \\ \mathrm{M} & \mathrm{M} & \mathrm{O} & \mathrm{M} \\ v_{n1} & v_{n2} & \Lambda & v_{nn_r} \end{bmatrix} \begin{matrix} t_1 \\ t_2 \\ \mathrm{M} \\ t_n \end{matrix} \\ r_1 \quad r_2 \quad \Lambda \quad r_{n_r} \end{matrix} \tag{12}$$

$$\tilde{Q} = Q \times T = \Phi(S) = \begin{matrix} \begin{bmatrix} v_1 & v_2 & \Lambda & v_n \end{bmatrix} \\ t_1 \quad t_2 \quad \Lambda \quad t_n \end{matrix} \tag{13}$$

$$\tilde{Q}_r = \tilde{Q}^T \circ \tilde{B}_r = \begin{matrix} \begin{bmatrix} v_1 & v_2 & \Lambda & v_n \end{bmatrix} \\ t_1 \quad t_2 \quad \Lambda \quad t_n \end{matrix} \circ \begin{matrix} \begin{bmatrix} v_{11} & v_{12} & \Lambda & v_{1n_r} \\ v_{21} & v_{22} & \Lambda & v_{2n_r} \\ \mathrm{M} & \mathrm{M} & \mathrm{O} & \mathrm{M} \\ v_{n1} & v_{n2} & \Lambda & v_{nn_r} \end{bmatrix} \begin{matrix} t_1 \\ t_2 \\ \mathrm{M} \\ t_n \end{matrix} \\ r_1 \quad r_2 \quad \Lambda \quad r_{n_r} \end{matrix} = \begin{matrix} \begin{bmatrix} v_1 & v_2 & \Lambda & v_{n_r} \end{bmatrix} \\ r_1 \quad r_2 \quad \Lambda \quad r_{n_r} \end{matrix} \tag{14}$$

$$\tilde{O} = \tilde{Q}_r @ \tilde{R}_r^T = \begin{matrix} \begin{bmatrix} v_1 & v_2 & \Lambda & v_{n_r} \end{bmatrix} \\ r_1 \quad r_2 \quad \Lambda \quad r_{n_r} \end{matrix} @ \begin{matrix} \begin{bmatrix} v_{11} & v_{12} & \Lambda & v_{1k} \\ v_{21} & v_{22} & \Lambda & v_{2k} \\ \mathrm{M} & \mathrm{M} & \mathrm{O} & \mathrm{M} \\ v_{n,1} & v_{n,2} & \Lambda & v_{n,k} \end{bmatrix} \begin{matrix} r_1 \\ r_2 \\ \mathrm{M} \\ r_{n_r} \end{matrix} \\ d_1 \quad d_2 \quad \Lambda \quad d_k \end{matrix} = \{d_1/v_1, d_2/v_2, \Lambda, d_k/v_k\} \tag{15}$$

$$\tilde{O}'_{\alpha} = \alpha\_cut(\tilde{O}, level) = \{d_1, d_2, \Lambda, d_{k_\alpha}\} \quad (16)$$

4. 두 모델 간 시간복잡도 비교

BK-퍼지정보검색모델은 내재되어 있는 검색 연산의 높은 시간복잡도를 가지고 그 적용분야를 축소한다. 문서의 개수가 n_D이고 용어의 개수를 n_T일 때, BK-퍼지정보검색모델의 퍼지검색요구의 시간복잡도는 $\Theta(n_D \times n_T)$이고 시소러스 구축의 시간복잡도는 $\Theta(n_D \times n_T^2)$이다.

개선된 BK-퍼지정보검색모델의 시간복잡도는 문서의 개수가 n_D, 용어의 개수를 n_T이고 축소용어가 원소 개수가 c_{n_r}이라고 할 때 문헌검색의 시간복잡도는 $n_D \times c_{n_r}$에 비례하고 시소러스를 구축의 시간복잡도는 $n_D \times n_T \times c_{n_r}$에 비례한다. 이때 c_{n_r}가 검색시스템과 검색결과의 신뢰도를 고려하여 산출된 상수임을 고려하면 문헌검색의 시간복잡도는 $\Theta(n_D)$이고 시소러스 구축의 시간복잡도는 $\Theta(n_D \times n_T)$임을 알 수 있다.

5. 개선된 BK-퍼지정보검색모델의 신뢰도 분석

문서검색시스템에서 검색결과의 신뢰도란 사용자의 질의어에 대한 검색된 문서의 적합도를 의미한다. 그러나 검색 문서의 적합도는 주관적인 것이고 비교 대상 역시 불확실하여 이를 정량적인 값으로 산정하는 것은 매우 어려운 일이다. 본 논문에서 제안하는 개선된 BK-퍼지정보검색모델의 검색 신뢰도 분석은 BK-퍼지정보검색모델의 검색결과와의 비교에 의한다.

5.1. 검색결과 신뢰도 측정

개선된 BK-퍼지정보검색모델의 검색 신뢰도는 BK-퍼지정보검색모델의 검색결과에 대한 유사도로써 산출하며 그 대상은 두 모델에 의한 검색결과에서 상위에 랭크되는 50개의 문서를 추출하여 구성한 두 개의 문서집합으로 한다.

본 논문에서는 두 개의 문서집합의 유사도를 산출하기 위해 Dice 상관계수(Dice coefficient)를 이용한다. Dice 상관계수는 벡터의 거리 및 유사도를 측정하는 방법 중의 하나로서 간소화와 정규화 기능을 가지고 있어 정보검색분야에 자주 쓰인다[8]. 두 개의 벡터 D_i, D_j가 주어질 때 Dice 상관계수 $S_{D_iD_j}$는 수식 (17)과 같다.

$$S_{D_iD_j} = \frac{2\sum_k weight_{ik} \cdot weight_{jk}}{\sum_k weight^2_{ik} + \sum_k weight^2_{jk}} \quad (17)$$

이때 문서집합은 이진 용어 가중치로 구성된 벡터로 볼 수 있다. D_i와 D_j는 두 개의 집합 A와 B로 표현가능하고 수식 (17)은 수식 (18)으로 바꿀 수 있다.

$$S_{AB} = \frac{2|A \cap B|}{|A| + |B|} \quad (18)$$

BK-퍼지정보검색모델의 상위 50개의 검색결과를 집합 α, 개선된 BK-퍼지정보검색모델의 상위 50개의 검색결과를 집합 β라 할 때 검색결과 α와 β의 유사도는 수식 (19)을 이용하여 산출 가능하다.

$$S_{\alpha\beta} = \frac{2|(\alpha \cap \beta)|}{50+50} = \frac{|(\alpha \cap \beta)|}{50} \quad (62)$$

5.2. 무작위 추출법에 의한 신뢰도 변화 유형

무작위 추출법은 난수를 이용하여 용어집합 중 일부를 축소용어집합으로 선정한다. 그림 1

은 무작위 추출법의 신뢰도의 변화 유형을 보여준다. 그 결과 축소용어집합의 크기가 작아질수록 그 신뢰도가 감소한다. 그리고 검색결과의 목표 신뢰도가 0.4일 때 무작위 추출법은 100 이내의 용어로 신뢰도를 만족시킬 수 있다.

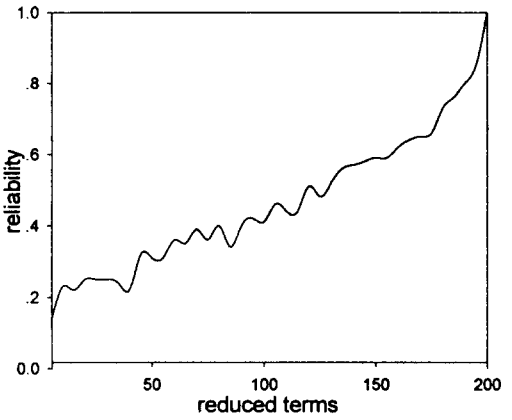

그림 1 무작위 추출법의 신뢰도 변화 유형

6. 결론

퍼지정보검색은 의료진단(medical diagnosis), 정보검색, 수기분류(handwriting classification) 등 수많은 문제 해결에 응용되고 있다[4]. 퍼지정보검색은 용어, 질의, 문헌과 같은 개체들의 관계를 퍼지관계행렬로 표현하고 개체연결과 같은 1차원적 단순검색 뿐만 아니라 개체간의 연관성에 근거하여 직접 관련이 없는 제3의 개체도 검색 가능하므로 단순 매칭(matching)으로 해결하기 힘든 화상, 동영상, 소프트웨어 재사용 등과 같은 분야에 특히 유용한 기법이다.

기존의 검색이론과는 달리 BK-퍼지정보검색모델만이 가지고 있는 특성에도 불구하고 실제 대용량의 문서나 용어를 다루는 검색시스템 적용이 어려웠던 것은 BK-퍼지정보검색모델 자체 검색연산의 높은 시간복잡도 때문이었다. 본 연구에서는 BK-퍼지정보검색모델의 시간복잡도를 개선한 개선된 BK-퍼지정보검색모델을 제안하고 이를 시간복잡도와 신뢰도 측면에서 분석하였다. 결론적으로 개선된 BK-퍼지정보검색모델은 시간복잡도 측면에서는 그 성능을 개선하였지만 검색 결과의 신뢰도 측면에서는 축소용어집합의 크기가 작아질수록 그 신뢰도가 감소함을 알 수 있다. 따라서 시스템 개발자는 시스템 처리능력과 검색결과의 목표 신뢰도 간 적정 타협점을 산정하여야 한다.

참고 문헌

[1] Frakes, William. B. and Baeza-Yates, Ricardo, Information Retrieval, Prentice Hall, 1992

[2] Keravnou, E. "Fuzzy Relational Products in Information Retrieval Systems." B. Tech. Dissertation, Dept. of Computer Science, Brunel University (1982)

[3] Kim, Yong-Gi and Kohout, L. J., "Use of Fuzzy Relational Products and Algorithms for generating Control strategies in resolution based Automated Reasoning." Proceedings of the fourth International Fuzzy System Association (IFSA) world congress, (Brussels, Belgium), July 7-12, 1991

[4] Kohout, L. J., and Harris, M., "Computer Representation of Fuzzy and Crisp Relations by Means of Threaded Trees Using Foresets and Aftersets." Journal of Fuzzy Logic and Intelligent Systems, vol. 3, no.1, 1993

[5] Kohout, L. J., Keravnou E. and Bandler W., "Automatic Documentary Information Retrieval by means of Fuzzy Relational Products." In Gaines, B. R., Zadeh L. A. and Zimmermann, H. J., editors Fuzzy

Sets in Decision Analysis, pages 308-404, North-Holland, Amsterdam, 1984

[6] Kohout, L. K., Bandler, W., "Fuzzy Relational Products as a Tool for Analysis and Synthesis of the Behaviour of Complex Natural and Artificial Systems." in: Wang S. K. and Chang P. P. eds., Fuzzy Sets: Theory and Application to Policy Analysis and Information Systems (Plenum Press, New York, 1980) 341-367.

[7] Rijsbergen, C. J. van, Information Retrieval, 2nd edition, butterworths, 1979

[8] Salton, G., Automatic Text Processing, Addison-Wesley, 1989

[9] Zimmermann, H. J. Fuzzy Set Theory and Its Application, Kluwer Academic Publishers, 1991