순수 신뢰도를 적용한 파이로 문서 분석

김성용, 박희성, 권은하, 김호동
한국원자력연구원, 대전시 유성구 대덕대로 989번길 111
sykim@kaeri.re.kr

1. 서론

연구 과정에서 발생되는 정보들은 전문가에 의해 의미 있는 지식으로 가공되고 문서화되어 저장소에 저장된다. 문제는 새로운 사용자가 접근하는 과정에서 문서에 저장된 내용들을 파악하고 습득하는데 시간과 비용이 많이 소모된다. 이에 대한 전문가가 내용을 전달하려 해도 직관에 의존하기 때문에 검토하지 못한 부분이 발생할 수 있으며, 이런 가능성은 지식의 양에 비례하여 높아진다. 이러한 문제점은 외부 협력 기관과 함께 업무를 추진할 때 공유해야할 내용 선정 및 검토 과정에서 비용 및 시간 소모, 내용 누락 등의 악영향을 초래하는 형태로 나타날 수 있다.

이러한 지식 관리에 대한 오류를 줄이는 일은 객관적으로 정보를 분석하여 활용함을 통해 이루어 질 수 있다.

본 연구에서는 한국원자력연구원에서 생성된 파이로 관련 문서들을 분석함에 있어 데이터 마이닝(data mining)의 연관 규칙 분석(association rule) 기법을 적용 하였다. 이에 대한 실험 결과와 분석 결론을 다룬다.

2. 본론

2.1 데이터 마이닝의 필요성 및 특징

정보 기술이 발전함에 따라 기업들의 데이터 베이스에는 방대한 양의 정보가 저장되었다. 이러한 데이터 베이스에서 쉽게 알 수 있는 정보들 외에도, 예측하지 못했고 쉽게 드러나지 않는 정보들까지 찾아내기 위해 데이터 마이닝이 사용된다.

데이터 마이닝의 여러 기법 중에 연관 규칙 분석이 있다. 연관 규칙 분석의 대표적인 활용 사례로 장바구니 분석이 꼽힌다. 상품의 매출과 관련된 데이터로부터 상품간의 연관성 정도가 큰 유형을 찾아 진열대의 상품 배치 및 홍보 등에 적용하여 매출을 증진시키는데 활용된다.

본 연구에서는 이와 같이 연관성에 대해 탐색하는 특성을 문서들에 저장되어 있는 키워드를 대상으로 사용했다.

2.2 순수 신뢰도의 적용

연관 규칙 분석은 항목 집합들 사이의 지지도(support), 신뢰도(confidence), 향상도(lift)등의 흥미도 척도(interestingness measure)를 바탕으로 연관성 정도를 측정한다[1]. 이 중에서 신뢰도는 가장 많이 활용되고 있고 있지만, 계산된 값만으로는 양의 연관성을 갖는지 음의 연관성을 갖는지 알 수 없어 음의 연관성까지 규칙으로 선택하게 되는 문제점을 가지고 있다. 본 연구에서는 이러한 연관성이 잘못 선택되는 오류를 피하기 위해 순수 신뢰도(net confidence)[2]를 적용하였다. 적용된 순수 신뢰도는 기존의 신뢰도만 적용한 실험에서[3] 알 수 없었던 새로운 척도를 나타낸다.

2.3 실험 결과

본 실험은 2011년 당시 한국원자력연구원 내 핵주기기술시스템공학부에서 생성된 문서들을 그 대상으로 하였다. 문서에서 한 개 이상 해당 키워드가 나올 경우 관련 키워드에 속한 한 개의 트랜잭션(transaction)으로 분류하였다.

Table 1과 Table 2는 미리 정의된 두 개의 키워드 설계(A), 요구사항(B)에 대한 7개 그룹의 순수 신뢰도를 측정한 결과이다.

Table 1. Confidence(A → B) & Net Confidence(A → B).

Group	Confidence (A → B)	Net Confidence (A → B)
Facility	0.0482	0.0473
Safeguards	0.0525	0.0519
Remote	0.1626	0.1626
Transport	0.0957	0.0857
System	0.0434	0.0414
Project	0.3795	0.3766
Manager	0.0891	0.0772
Total	0.1121	0.1111

Table 2. Confidence(B → A) & Net Confidence(B → A)

Group	Confidence (B → A)	Net Confidence (B → A)
Facility	0.9148	0.7577
Safeguards	0.8091	0.7684
Remote	0.9867	0.9711
Transport	0.8787	0.4716
System	0.3333	0.3112
Project	0.9113	0.8647
Manager	0.8095	0.4655
Total	0.8890	0.8326

신뢰도는 선행 항목이 포함된 트랜잭션에서 후행 항목이 함께 나타나는 비율이다. 순수 신뢰도는 선행 항목이 포함된 트랜잭션에서 후행 항목이 함께 발견되는 현상이 선행 항목을 포함하는 트랜잭션들만이 가지는 고유한 특성인지를 나타내는 수치이다. 순수 신뢰도는 [-1.1]의 범위를 가진다.

Table 1과 Table 2에서 전체적으로 A → B 보다 B → A 경우가 높은 수치를 보이는데 이는 실제로 A라는 상위 개념 안에 B라는 하위 개념이 속하는 실제 데이터들의 특징이 나타난 것이다. 그리고 양쪽 모두 순수 신뢰도가 양수 값이 나왔으므로 사용자가 정의하는 최저 순수 신뢰도보다 높은 관계는 흥미로운 규칙으로 분류 될 수 있다. 이렇게 도출된 흥미로운 규칙은 관계와 해당 수치가 정량적으로 존재함으로 사용자가 데이터들의 특성을 파악하는데 유용하다.

3. 결론

본 연구에서는 협조 부서 및 협력 업체와의 업무 추진부터 효율적인 지식 분석까지 응용하기 위해 데이터 마이닝을 파이로 관련 문서에 적용하여 정량적인 분석 결과 도출 가능성을 검증했다. 그 결과 실제 적용 가능한 긍정적인 수치가 발생하였고 이를 바탕으로 차후 개선하여 자동화된 분석 프로세스를 구축한다면 연구 발전에 이바지 할 것으로 기대된다.

4. 감사의 글

본 연구는 교육과학기술부의 원자력 연구개발사업의 일환으로 수행되었으며, 이에 감사드립니다.

5. 참고문헌

[1] Pang-Ning Tan, Michael Steinbach, Vipin Kumar, Introduction to Data Mining, Pearson Education; ISBN 0-321-42052-7, pp 327-486.

[2] 안광일, 김성집, 연관규칙 탐색에서의 새로운 흥미도 척도의 제안, 대한산업공학회지, 29, 41-48, 2003.

[3] 김성용, 최효연, 박희성, 최영, 김호동, 연관규칙을 적용한 요구사항 분석, 한국 소프트웨어공학 학술대회, 2012.

[4] Berzal. F, Blanco. L, Sanchez. D and Vila. M.A, A New Framework to Assess Association Rules, Proc. 4th Int. Conf. On Intelligent Data Analysis, 95-104, 2001.

[5] 박희창, 연관 규칙 마이닝에서 기여 순수 신뢰도의 제안, 한국데이터정보과학회지,22(2), 235-243, 2011.

[6] Tan. P.N, Kumar. V and Srivastava. J, Selecting the right interestingness measure for association patterns, Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 32-41, 2002.