효율적인 문서 처리 작업을 위한 문서집합 나열

나종열○ 문일철 최기선
KAIST 전자전산학과 전산학전공
{cynah, imoon, kschoi}@world.kaist.ac.kr

Sequencing Document Clusters to Support Human Annotation Efforts

Chong-Yeol Nah○ Il-Chul Moon Key-Sun Choi
Division of Computer Science, KAIST

요 약

온톨로지 어노테이션(Annotation)과정은 수동에 의해 대부분의 문서 처리 작업이 진행되고 있다. 그러므로 계획적이지 않은 문서 처리는 자칫 일관성이나 효율성을 떨어뜨릴 수가 있다. 예를 들어, 처리되는 문서들의 도메인이 자주 바뀌면, 수동 어노테이션을 하는 작업자는 객관성을 잃을 가능성이 높다. 따라서, 본 연구에서는 처리되는 문서집합의 도메인이 최대한 연관성이 유지되도록 처리 문서를 집합하여 나열하는 알고리즘을 소개한다. 첫번째로, 문서들을 유사한 주제 집합으로 형성한다. 두번째로, 두개 이상의 집합에 겹치는 문서들을 계산한다. 마지막으로, 이러한 겹침이 최대화 되도록 문서들의 처리 순서를 나열한다. 본 알고리즘을 IT관련 위키피디아 문서집합을 이용하여 평가를 시행했다. 평가 결과 우리의 알고리즘을 이용하면 처리되는 문서들의 도메인 이동이 무작위로 처리하는 것 보다 연속적이었음을 수치상으로 계산할 수 있었다.

1. 서 론

온톨로지는 복잡화된 현대의 지식, 정보, 관습, 체계를 표현하는데 있어서, 가장 적절한 방법으로 널리 사용되고 있다 [1]. 예를 들어, 자동차 정비의 복잡한 일을 하기 위해서, 자동차를 이루는 부품의 온톨로지를 구축하고, 이 온톨로지를 여타의 추론에 이용하여 복잡한 정비를 지원할 수 있다. 이런 온톨로지의 구축은 (반)자동화된 프로그램을 이용하거나 [2] 사람이 직접 작성할 수도 있다 [3]. 이런 추세 중에서도 사람이 만든 온톨로지는 그 정확도나 자세함에 있어서, 기계로 만든 온톨로지보다 우위에 있으며, 더 널리 사용되고 있다. 그러나 사람이 만든 온톨로지는 기계로 만든 온톨로지보다 일관성의 유지가 힘들고, 전체적인 평가를 하며 작성하기 힘들고, 많은 양을 만들기도 힘들다. 그러므로 우리는 인간의 온톨로지 작성을 도와주기 위한 문제 해결 방법이 필요하다. 특히 온톨로지 작성의 일관성을 유지하는 면에 있어서, 온톨로지 작성의 배경이 되는 처리 문서의 주제가 자주 바뀌면, 인간은 일관성을 쉽게 잃게 된다 [4]. 그러므로 우리는 이 연구를 통해서, 주제가 최대한 바뀌지 않으면서 일관성있게 인간이 작업할 수 있도록 배경 문서를 나열해 주는 문제 해결 방법을 소개한다.

우리의 문제 해결 방법은 우선 문서를 집합화하여 주제 단위로 묶고, 묶여진 문서집합들을 가장 영향력이 큰 순서대로 나열하며, 그리고 그 세부 문서들을 가장 주제 연관성이 높도록 다시 나열 하는 방식이다. 우리는 이 문제 해결 방법을 구현하여, 위키피디아의 IT관련 문서들을 이용하여 시험하였다. 시험결과는 우리의 문제 해결 방법이 무작위로 문서를 나열하는 방법보다, 더욱 주제를 일관성 있게 제시하는데 효과적이라는 결과를 보여주었다. 이 문제 해결 방법은 인간이 문서를 이용하여 작업하는 여러 컴퓨터 작업 환경에 적용될 수 있을 것이다. 왜냐하면 인간이 문서를 작업할 때 일관성을 유지하기 힘들다는 것은 종종 발견되는 사실이기 때문이다 [4]. 그러므로 이 문제해결 방식은 일반적인 자연언어 문서처리의 여러 시스템에 적용될 수 있을 것이다.

2. 관련연구

온톨로지 제작 과정에서의 자동 및 수동 방법의 장단점을 살펴보고 이를 돕기 위해 문서 집합을 군집화 하여 나열하는 방법들에 대해 설명하겠다.

2.1. 온톨로지의 구축에 대한 인간의 노력

Text2Onto [2] 프로그램은 문서 집합에서 개념간의 관계를 자동으로 추출하는 툴이다. 이 시스템의 특징은 한 문서 집합에서 개념들 간의 상하위 체계와 인스턴스를 자동으로 추출하는 장점이 있으나 추출된 정보의 검증이 안 된다는 단점이 있다.

Stanford Univ. KSL에서 제작한 Ontolingua [3] 시스템은 수동 온톨로지 제작을 도와주는 툴이다. 이 프로그램의 특징은 사용자로 하여금 온톨로지 제작 환경을 제공함으로서 개념 또는 개념간의 관계들을 편리하게 검

색, 제작, 수정 또는 재이용을 가능하게 해주는 것이다. 사용자 편의를 고려한 인터페이스를 제공은 하지만 모든 과정이 반수동으로 이루어지므로 생산성이 낮은 단점이 있다.

2.2. 온톨로지의 구축 배경문서를 나열하는 방법

Kleinberg의 [5] "Hubs and Authorities" 알고리즘은 문서간의 링크 정보를 이용하여 비슷한 문서끼리 집합을 형성하는 것이다. 문서들의 링크구조를 이용하여 각 문서를 hub page와 authorities page로 구분하고 정해진 hub page를 이용하여 문서들을 집합화 하는 방법을 소개하고 있다.

Lin et al.은 [6] MEDLINE 문서들에 대한 지식추출 정확도를 높이기 위하여 문서들을 주제별로 군집화 하고 각 문서집합의 중요도에 의해 우선순위를 부여하는 알고리즘을 연구했다. MEDLINE 문서들의 citation정보에서 키워드를 사용하여 문서들을 집합화 하고 각 문서의 citation count per year (CCPY), citation count (CC), and journal impact factor (JIF)을 이용하여 각 집합의 중요도를 결정하였다.

본 논문에서는 위키피디아의 "List_of_XXX" 페이지를 이용하여 문서들을 집합화 하고 문서중복이 최대화가 되도록 문서집합들을 나열하여 그림1과 같이 사람의 문서처리 작업 효율을 높이는 것이 목적이다.

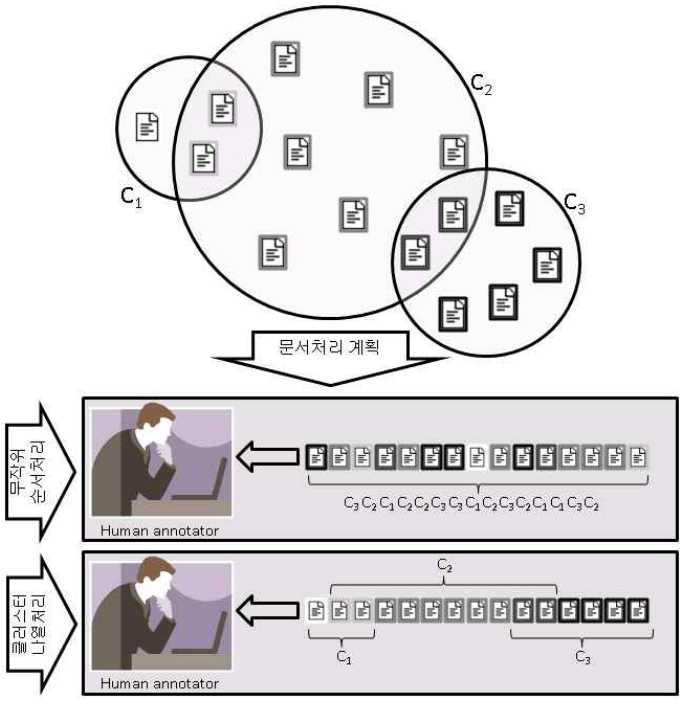

그림 100 문서집합 나열을 통한 문서처리 계획

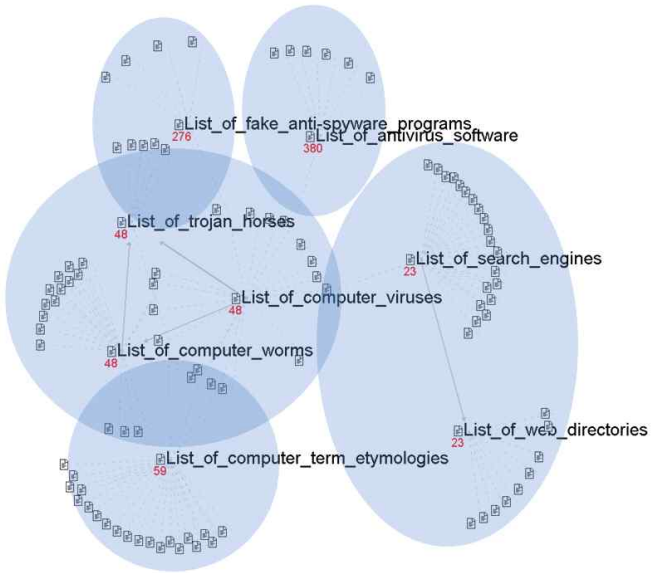

그림 101 문서집합들과 집합겹침 현상

3. 방 법

문서 처리의 일관성을 유지하는 방법으로, 우리는 1) 문서를 주제에 따라 집합화하고, 2) 만들어진 문서 집합들의 겹침을 최대화하도록 문서를 나열하는 알고리즘을 선택하였다. 그러므로 우리는 이 장을 통해, 문서의 집합화 방법과 겹침을 최대화하는 나열방법에 대해 소개한다.

3.1 문서 집합 생성 방법

문서의 집합은 그 안에 속한 문서들이 같은 도메인이 되도록 만들었다. 위키피디아 문서들을 집합화할 때 카테고리 정보를 사용하지 않은 이유는 그 카테고리 분류는 사람의 임의대로 한 것이므로 문서의 분류가 자의적인 케이스가 많다. 따라서, 본 연구에서는 'List_of_XXX' 제목의 hub page들을 가지고 문서들의 집합들을 형성 하였다. 위키피디아 카테고리 정보는 하나의 문서가 여러 도메인에 포함이 되어야할 경우 정보가 결여되거나 부정확할 수 있지만 'List_of_XXX' 페이지 안의 문서들은 해당 도메인을 중심으로 그 안에 해당되는 모든 페이지를 모은 것이기 때문에 페이지간의 연관성이 더 높다. Hub page들은 또한 그들끼리의 링크를 갖고 있기 때문에 우선 hub page간의 링크를 모두 찾아서 401개의 허브페이지 집합을 발견했고, 각 hub page 집합의 hub page에 링크되어 있는 페이지들을 찾아 하나의 문서 집합을 형성했다 (그림2).

3.2 문서의 중복을 최대화하기 위한 알고리즘

하나의 문서는 일반적으로 한개 이상의 hub page에 속하게 된다. 예를 들어 'Database_algorithm' 페이지는 'List_of_Algorithms'라는 허브페이지와 'List_of_Database_Algorithms' 허브페이지에 모두 링크가 있다. 문서처리를 하는 작업자의 입장에서는 이러한 중복되는 페이지들이 많은 집합들부터 일처리를 하면 이후의 작업 부담이 덜어지게 되기 때문에 클러스터 시퀀스를 중복페이지가 많은 클러스터부터 나열하는 알고리즘을 만들었다.

우선 Page distribution(*pd*)과 Cluster priority(*cp*) 두 개의 변수를 설정하였다. Page distribution(*pd*)은 해당 페이지가 포함된 문서집합의 개수이다. 즉, *pd*값이 높으면 그만큼 그 페이지는 여러 도메인에서 자주 나타나는 페이지라는 의미이다. Cluster priority(*cp*)는 해당 문서집합에 포함된 페이지들의 *pd*값의 총합이므로 *cp*값이 높을수록 빈도수가 높은 페이지들이 많은 문서집합이다. Page distribution과 Cluster priority 값들을 구하는 공식은 표1과 같다.

표 1 Page distribution과 Cluster priority 공식

$$P = \{p_1, p_2, ..., p_M\}, \text{N은 총 페이지개수}$$

$$C = \{c_1, c_2, ..., c_M\}, \text{M은 총 클러스터개수}$$

$$involve(p_i, c_j) = \begin{cases} 1 & \text{if } p_i \in c_j \\ 0 & \text{else} \end{cases}$$

$$pd(p_i) = \sum_{c_j \in C} involve(p_i, c_j)$$

$$cp(c_j) = \sum_{p_i \in c_j} pd(p_i)$$

4. 평 가

실험에 사용된 문서들은 7532개 위키피디아 IT관련 페이지를 사용했다. 이 페이지들은 IT온톨로지 제작을 위해 작년 CoreOnto 프로젝트에서 사용한 코퍼스이다.

본 연구에 대한 평가는 처리되는 문서 개수에 대한 완료되는 문서집합의 개수의 비율로 하였다. 완성되는 문서집합의 개수가 빠를수록 처리되는 문서들의 도메인이 자주 바뀌지 않는다는 것을 뜻한다.

4.1 클러스터링 결과

전체 7532개 페이지 중 hub page는 568개였으며 총 401개의 hub page 집합을 형성했다. 하나의 hub page 집합은 1개에서부터 19개까지의 hub page로 구성되는 집합들이 생성되었다. 각 hub page 집합에서 문서 집합을 만든 결과 1개에서 부터 1086개의 문서를 지니고 있는 문서 집합들이 생성되었다. 문서가 1개인 문서집합은 집합 안에 hub page만 존재한다는 뜻이므로 이러한 집합들은 관심 대상에서 제외하였다. 총 401개의 문서집합 중 hub page만 포함되어있는 집합은 38개이므로 문서집합 나열에 적용되는 문서 집합은 총 363개이다.

표 2 문서 집합 세부 통계 수치

항 목	수 치
총 페이지 개수	7532
총 wikipedia page 개수	6964
총 hub page 개수	568
총 클러스터 개수	401
hub page만 있는 집합 개수	38
관심대상 클러스터 개수	363

4.2 문서집합 나열

문서집합 나열의 효용성을 알아보기 위해 베이스라인을 무작위 문서 선택으로 정했다. 랜덤 시드값 1에서부터 30까지를 사용하여 실험한 30개의 실험 평균과 문서집합 나열을 통한 실험을 시행하여 그 성능을 비교했다.

실험결과 문서집합 나열을 이용하여 작업을 한 결과 그림3과 같이 작업 완료되는 문서집합의 숫자가 월등히 빠름을 확인할 수 있다. 이것은 문서처리의 효율이 더욱 우수하다는 뜻이며 또한 하나의 문서집합 단위로 일을 처리하기 때문에 작업자의 입장에서 문서들의 도메인 변동이 많지 않음을 나타낸다.

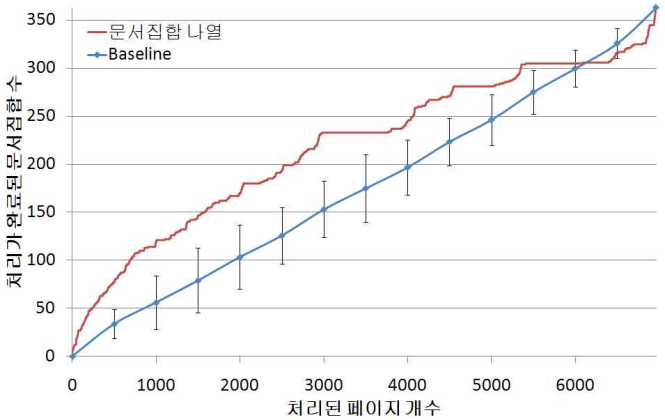

그림 102 문서처리 계획에 따른 문서집합 완료 속도
(Baseline은 문서를 30번 무작위로 나열한 결과)

5. 결 론

우리는 문서를 집합화하여 주제 단위로 묶고, 묶여진 문서 집합을 가장 영향력이 큰 순서대로 나열하며, 그 세부 문서들을 가장 주제 연관성이 높도록 재나열 하는 방식을 사용했다.

우리는 이 문제 해결 방법을 구현하여, 위키피디아의 IT관련 문서들을 이용하여 시험하였다. 시험결과는 우리의 문제 해결 방법이 무작위로 문서를 나열하는 방법보다, 주제를 일관성 있게 제시하는데 효율적이라는 결과를 보여주었다.

무작위로 나열한 문서순서로 작업을 한 결과 처리된 문서들이 많아도 완성된 문서집합의 개수가 증가하지 않았다. 그 뜻은 문서를 처리할 때 하나의 도메인에 집중을 하지 않고 산재되어 있는 문서집합에서 작업하고 있음을 알 수 있다.

반대로 문서집합 나열을 하여 문서처리를 한 결과 완성되어지는 문서집합의 개수가 빠르게 증가했으며 그 이유는 도메인단위로 문서처리를 하기 때문에 도메인 이동이 적었음을 뜻한다.

이 문제 해결 방법은 인간이 문서를 이용하여 작업하는 여러 컴퓨터 작업 환경에 적용될 수 있을 것이다. 왜냐하면 인간이 문서를 작업할 때 일관성을 유지하기 힘들다는 것은 종종 발견되는 사실이기 때문이다 [4]. 그러므로 이 문제해결 방식은 일반적인 자연언어 문서 처리의 여러 시스템에 적용될 수 있을 것이다.

Acknowledgements

우리의 연구는 정보통신연구진흥원의 지원 (A1100-0601-0102)을 통해 이루어졌다. 이 논문의 내용은 오로지 저자의 견해로 보아야하며, 지원기관의 의견을 나타내지 않는다.

이 논문은 2008년도 한국과학기술원 BK21 정보기술사업단에 의하여 지원되었습니다.

References

[1] Golebiowska J (2000) SAMOVAR - Setting up and Exploitation of Ontologies for Capitalising on Vehicle Project Knowledge, Proceedings of EKAW'00, Workshop on Ontologies and Texts, Juan-les-Pins

[2] Cimiano, P. and Volker, J. (2005) Text2Onto A Framework for Ontology Learning and Data-driven Change Discovery. Proceedings of NLDB, pp. 227-238

[3] Knowledge Systems Laboratory (KSL), Stanford University. "Ontolingua".
http://www.ksl.stanford.edu/software/ontolingua/

[4] Proulx, J. (2004) The Enactivist Theory of Cognition and Behaviorism: An Account of the Processes of Individual Sense-Making ROULX, Proceedings of the Complexity Science and Educational Research Conference, pp. 115-120

[5] Kleinberg, J. (1998) Authoritative sources in a hyperlinked environment, Proceedings of ACM-SIAM Symposium on Discrete Algorithms

[6] Lin, Y., Li, W., Chen, K., and Liu, Y. (2007) A Document Clustering and Ranking System for Exploring MEDLINE Citations, Journal of the American Medical Informatics Association, Vol. 14, Issue. 5, pp. 651-661