creative commons

COMMONS DEED

저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:

저작자표시. 귀하는 원저작자를 표시하여야 합니다.

비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.

변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 이용허락규약(Legal Code)을 이해하기 쉽게 요약한 것입니다.

Disclaimer

dCollection

조선대학교
CHOSUN UNIVERSITY

[UCI]I804:24011-200000267330

2019년 2월

박사학위 논문

학술지 추천을 위한 WordNet 기반 카테고리 유사도 측정

조선대학교 대학원

컴퓨터공학과

고 병 규

조선대학교
CHOSUN UNIVERSITY

학술지 추천을 위한 WordNet 기반 카테고리 유사도 측정

Measuring WordNet-based Category Similarity for Academic Journal Recommendation

2019년 2월 25일

조선대학교 대학원

컴퓨터공학과

고 병 규

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

학술지 추천을 위한 WordNet 기반 카테고리 유사도 측정

지도교수 김 판 구

이 논문을 공학박사학위 신청 논문으로 제출함

2018년 10월

조선대학교 대학원

컴퓨터공학과

고 병 규

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

고병규의 박사학위논문을 인준함

위원장	조선대학교	교수	정 일 용	(인)
위 원	조선대학교	교수	양 희 덕	(인)
위 원	숭실대학교	교수	홍 지 만	(인)
위 원	중앙대학교	교수	정 재 은	(인)
위 원	조선대학교	교수	김 판 구	(인)

2018년 12월

조선대학교 대학원

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

목 차

Ⅰ. 서론 ····· 1
A. 연구 배경 및 목적 ····· 1
B. 연구 내용 ····· 2
C. 논문 구성 ····· 4

Ⅱ. 관련연구 ····· 5
A. 학술정보서비스 소개 및 현황 ····· 5
1. 네이버 학술정보 ····· 5
2. 구글 스칼라 ····· 6
3. Microsoft Academic Search ····· 7
4. SCOPUS ····· 7
B. 학술지 평가에 사용되는 지표 ····· 8
1. Impact Factor ····· 8
2. Eigenfactor Score ····· 8
C. WordNet ····· 9

Ⅲ. 학술지 카테고리 분류 및 연구 실적 데이터 분석 방법 ····· 10
A. 학술지 카테고리 분류 방법 ····· 10
1. JCR 학술지 카테고리 수집 및 분석 방법 ····· 10
2. 카테고리 내 학술지 간 관계 정의 ····· 13
B. Google Scholar를 이용한 연구 실적 수집 방법 ····· 15
1. HTML문서 전처리 방법 및 정보 추출 방법 ····· 15
2. 연구 실적 분석결과 표현 방법 ····· 22

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

Ⅳ. 카테고리 유사도 측정 방법 ····· 24
A. 거리 측정 알고리즘 ····· 24
1. 맨해튼 거리 ····· 24
2. 유클리디안 거리 ····· 25
3. 민코프스키 거리 ····· 26
B. 거리 측정을 통한 학술지 간 거리 측정 방법 ····· 27
C. EigenValue 기반 카테고리 유사도 측정 방법 ····· 31
1. 카테고리명을 구성하고 있는 단어 간 유사도 측정 ····· 31
2. 각 학술지별 논문 주제(Subject, Abstract, Indexing)를 활용한 유사도 측정 ····· 32
3. 카테고리 유사도 측정 방법 ····· 34
D. 카테고리 유사도를 이용한 학술지 추천 ····· 38

Ⅴ. 결론 및 향후 연구 ····· 43
A. 결론 ····· 43
B. 향후 연구 ····· 44

참고문헌 ····· 45

부 록 ····· 48

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

표 목 차

[표 3-1] JCR에 정의된 저널 카테고리 ····· 11
[표 3-2] ‘Computer Science, Information Systems’ 카테고리 정보 ····· 12
[표 3-3] ‘Computer Science, Information Systems’ 내 포함된 학술지 정보 ····· 12
[표 3-4] 두 개의 카테고리 사이에 동시에 포함되는 저널의 수 ····· 15
[표 3-5] Google Scholar 내 개인연구실적 정보 추출 단계 ····· 16
[표 3-6] HTML Parser를 사용하기 위한 추출 규칙 정의 ····· 19
[표 3-7] Google Scholar 프로필 내 주요 데이터 추출 과정 ····· 20
[표 3-8] 학술지별 빈도수 측정 결과 및 학술지 지표 ····· 22
[표 4-1] 학술지별 빈도수 측정 결과 및 학술지 지표 ····· 27
[표 4-2] Wu & Palmer Similarity를 활용한 단어 간 유사도 측정 결과 ····· 32
[표 4-3] ‘IEEE ACCESS’와 ‘Neurocomputing’ 학술지 내 다루고 있는 학술분야 ····· 33
[표 4-4] ‘IEEE ACCESS’와 ‘Neurocomputing’ 학술지에 포함된 초록을 구성하는 단어 간 유사도 측정결과 ····· 34
[표 4-5] EigenValue(λ)값 계산 결과 ····· 35
[표 4-6] 평균값을 구하기 EigenValue 계산 결과 예제 ····· 36
[표 4-7] Lamda A, B에 대한 각각의 평균값 ····· 36
[표 4-8] 저자별 연구 결과물 현황 ····· 38
[표 4-9] ‘Computer Science, Information Systems’와 ‘Computer Science, Software Engineering’ 간 특성값 추출 결과 ····· 40
[표 4-10] Author A의 연구 실적 중 “Multimedia Tools and Applications”를 기준으로 “Computer Science, Information Systems” 카테고리 내 학술지와의 거리값 측정 결과 ·40
[표 4-11] Author A의 연구 실적 중 “Multimedia Tools and Applications”를 기준으로 “Computer Science, Software Engineering” 카테고리 내 학술지와의 거리값 측정 결과 41
[표 4-12] Author A에 대한 학술지 추천 목록 ····· 41

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

그 림 목 차

[그림 2-1] 연도별 “Machine Learning” 대한 연구 트렌드 검색 결과 ·····5
[그림 3-1] 데이터 전처리 및 Keyword Weight 산출을 위한 전체 구성도 ·····10
[그림 3-2] JCR 내 ‘BURNS’ 학술지의 카테고리 정보 ·····13
[그림 3-3] 각 카테고리별 학술지 계층 구조 및 학술지 간 관계 형성 예제 ·····14
[그림 3-4] Google Scholar에서 사용하는 프로필 검색 쿼리문 ·····16
[그림 3-5] HTML Parser에서 사용할 최종 URL ·····17
[그림 3-6] Google Scholar에서 제공하는 개인 프로필 정보 ·····17
[그림 3-7] 개인 프로필 정보 웹 문서의 HTML 소스 정보 ·····18
[그림 3-8] 개인 프로필 정보 웹 문서의 HTML 소스 정보 ·····18
[그림 3-9] Google Scholar 내 연구 실적을 기반으로 한 학술지 분포도 ·····23
[그림 4-1] 맨해튼 거리 알고리즘 예 ·····24
[그림 4-2] Google Scholar 내 연구 실적을 기반으로 한 저널 분포도 ·····28
[그림 4-3] Google Scholar 내 연구 실적을 기반으로 한 저널 분포도 ·····29
[그림 4-4] 특정 학술지 AB-1이 속해있는 두 카테고리 내 학술지 간 거리 측정 예 ·····30
[그림 4-5] WordNet 기반 ‘Computer’ 단어에 대한 계층구조 ·····31
[그림 4-6] 기하평균을 사용한 최적의 좌표값을 구하는 방법에 대한 예 ·····37
[그림 4-7] Author A의 연구실적 결과를 나타낸 그래프 ·····39

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

ABSTRACT

Measuring WordNet-based Category Similarity for Academic Journal Recommendation

Byeongkyu Ko
Advisor : Prof. Pankoo Kim, Ph.D
Department of Computer Engineering
Graduate School of Chosun University

With research activities based on the new computing environment in diverse areas, the systems for analyzing and verifying the researchers' research competencies have appeared. Google Scholar provided by Google automatically collects/stores personal profiles to verify each researcher's research competency index and individual outcomes at a time. Scopus provided by Elsevier provides diverse information by collecting theses, books, and journals published by relevant institutes, and then storing information. Even though the services for the provision/analysis of research competency information for researchers have been repetitively and continuously advanced, the methods for users to easily find necessary information are not efficiently provided.

Especially, the existing systems perform the classification based on the preceding research area, and a single research area category includes one or as many as 300 journals. Thus, each journal was not classified into a single category, but included in multiple categories. Like this, the collected/classified journals provide information to users through diversely-converted indexes based on each information(the number of quotation, the number of being quoted, and etc). As this provision of research competency information is mostly used for the evaluation of journals, instead of

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

individuals, however, there are lots of difficulties for users to use it. If the overall system is not fully understood, it is hard to get sufficient information from it. In the currently-provided research competency analysis service, it is also hard to verify the research area of users and the information for the development of research competencies. In order to re-verify the research area of researchers and also to develop the research competencies based on it, the information about journals was collected and analyzed, and the information of each category diversely classified was also collected.

In this process, the relationship between journals could be expressed through the impact factors and eigenvector scores generally used out of the journal evaluation indexes. However, there were lots of difficulties to find relationships between categories, so that this thesis aims to suggest a method to solve this. Therefore, this thesis aims to provide the information of category that could improve the researchers' research abilities and clearly show their research areas, by easily showing their research areas through information related to other journals based on their own research outcomes, and also recommending the relevant proper journals.

For this, the measurement of similarity between categories based on the category information provided by the existing academic information service was suggested. In case of the journals included in two categories or more, the distance between journals was measured based on the information included in each category, and then the measurement of similarity between categories suggested by this thesis was applied. And, total five optimum journals were recommended.

In the results, it became possible to recommend journals in wider scope than the recommendation of journals within a category. Also, by suggesting the clear grounds for the method of measuring the similarity between categories, the ambiguity related to the process for the measurement of similarity could be resolved. It is meaningful in the aspect of suggesting the method that could be expansively applied to the user customized-service recommendation area, on top of recommendation of journals.

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

Ⅰ. 서론

A. 연구배경 및 목적

새로운 컴퓨팅 환경과 이를 기반으로 다양한 분야에서 연구 활동이 진행됨에 따라 연구자들의 연구 역량을 분석하고 이를 확인할 수 있는 시스템들이 출현하였다. Google사에서 서비스 중인 Google Scholar는 연구자들의 개인 연구 역량 지표 및 개인의 결과물을 한번에 확인할 수 있도록 개인 프로필을 자동으로 수집 저장한다. Elsevier사에서 서비스 중인 Scopus는 해당 기관에서 출판하는 논문, 도서, 저널등을 수집하여 정보를 저장하고, 이를 바탕으로 다양한 정보를 제공한다. 이처럼 연구자들에게 제공하고자 하는 연구 역량 정보 제공 및 분석 서비스는 날로 발전을 거듭하고 있으나, 사용자들이 필요한 정보를 쉽게 찾는 방법을 효율적으로 제공하지 못하고 있다. 특히, 기존의 시스템들은 먼저 연구분야를 기반으로 분류하고 있으며, 하나의 연구분야 카테고리 내 한 개에서부터 많게는 300가지가 넘는 학술지를 포함하고 있다. 이로 인해 각각의 학술지는 한 개의 카테고리에 분류되지 않고 다수의 카테고리에 포함되어 있다는 부분까지 확인 할 수 있었다. 이와 같이, 수집 및 분류된 학술지들은 각각의 정보(인용 수, 피인용 수 등)를 바탕으로 다양한 지표로 환산하여 사용자에게 정보를 제공하지만, 이러한 연구 역량 정보 제공은 대부분 사용되는 범위가 사용자 개인이 아닌 학술지에 대한 평가를 위해 사용되기 때문에 사용자가 사용하기에는 많은 어려움이 발생하고, 이를 사용할 때 전반적인 시스템을 이해하지 않고 사용한다면 충분한 정보를 얻기가 어렵다. 또한 사용자의 연구분야를 확인하고, 연구 역량을 발전시키기 위한 정보를 확인하는 것도 현재까지 서비스 되고 있는 연구 역량 분석 서비스에서는 찾아보기 힘들다.

이러한 연구자들의 연구 분야를 재확인하고 이를 기반으로 한 연구 역량을 발전시키기 위해 학술지에 대한 정보를 수집하여 분석하였으며, 다양하게 분류되어 있는 각각의 카테고리 정보도 수집을 할 수 있었다. 이 과정에서 학술지 평가 지표 중 일반적으로 사용되는 Impact Factor와 Eigenvector Score를 통해 학술지 간 관계를 표현할 수 있었으나, 카테고리 간의 관계를 찾는데 기준 및 관계 선정이 모호함에 따라 문제를 해결하기 위한 다양한 연구가 진행되었던 사실을 확인하고 본 연구를 통해 방법을 제안하고자 한다.[1, 2].

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

따라서, 본 논문에서는 자신의 연구 결과물을 바탕으로 타 학술지와의 연관정보를 통해 자신의 연구 분야를 쉽게 확인하고, 이와 관련한 적절한 학술지를 추천해 줌으로써 연구자들의 연구 능력을 고취시키고 연구자의 연구 분야를 명확히 판단할 수 있는 카테고리 정보를 제공하고자 한다. 이를 위해 기존 학술 정보 서비스에서 제공하는 카테고리 정보를 바탕으로 카테고리 간의 유사도 측정 방법을 제안하며, 두 개 이상의 카테고리에 포함되어 있는 학술지는 각각의 카테고리 내 포함된 정보를 바탕으로 학술지 간의 거리를 측정하고 본 논문에서 제안하는 카테고리 간 유사도 측정 방법을 적용한 후 총 5개의 최적합 학술지를 추천하였다.

그 결과 한 카테고리 내에서 학술지를 추천했을 때보다 더 넓은 범위의 학술지 추천이 가능하였다. 또한, 카테고리 간 유사도 측정 방법에 대한 명확한 근거를 제시함으로써 유사도 측정을 위한 과정과 관련한 모호성을 해소할 수 있었다. 이를 바탕으로 학술지 추천뿐만 아니라 사용자 맞춤형 서비스 추천 분야에도 확장 적용할 수 있는 방법을 제안했다는 점에서 본 연구의 가치를 확인할 수 있었다.

B. 연구 내용

본 연구에서는 기존에 제공되는 각 학술지별 상위 카테고리를 분석하여 카테고리 내 존재하는 학술지 상하위 관계를 정리한다. 그리고 Google Scholar에서 특정 인물이 제출 후 게재된 학술지 목록 및 논문 제목 등을 추출하여 분석 후 카테고리 내 학술지순위와 Impact Factor를 사용하여 2차원 평면좌표에 표현한 후 연구자들이 자신의 연구 분야를 확인할 수 있도록 표현한다. 또한, 카테고리 유사도 측정 방법을 적용하여 최적합 학술지를 추천한다. 위의 연구를 수행하기 위하여 아래와 같은 연구 과정을 거치게 된다.

1. 학술지 카테고리를 정의

기존 학술정보서비스 내 분류된 카테고리 정보를 수집하여 학술지와 카테고리 간 상하위 관계를 정의하고 학술지의 중복포함 관계를 분석 및 정의한다. 또한, 거리 측정 방법을 사용하여 카테고리 내 특정 학술지와의 거리를 측정함으로써 학술지 간 유사도 측정 방법을 제시한다.

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

2. 사용자의 연구 결과물 수집 방법 및 전처리

사용자의 연구 결과물 수집을 위해 Google Scholar 내 사용자의 연구 실적 데이터 수집 방법을 설명하고, 이를 위해 HTML Parser를 사용하여 HTML 문서 전처리 방법을 설명한다. 문서 전처리 후 최종 추출된 값을 사용하여, 연구 실적 데이터와 학술지 데이터 간 분석을 통해 본 연구에서 요구하는 추가 정보와 매칭하여 데이터 베이스에 저장한다.

3. 카테고리 유사도 측정 방법 제안

카테고리 유사도 측정을 위하여 지식베이스를 활용한 단어 간 유사도 측정 방법을 사용한다. 이는 카테고리명을 선정할 때, 해당 카테고리를 대표하는 단어들을 사용하기 때문에 카테고리를 이루고 있는 단어들을 대표어로 선정하였기 때문이다. 또한, 카테고리 내 포함된 각 학술지에서 다루는 논문의 카테고리를 추가로 활용하여 각 각각의 모집분야에 대한 목표 단어들을 추출하여 앞서 선정한 대표어와 함께 유사도 측정을 한다. 이를 바탕으로 Eigenvalue 방법을 이용한 카테고리 유사도 측정 방법 제안한다.

4. 2차원 평면그래프를 통해 현재 연구분야를 제시 후 5개의 최적합 학술지를 추천

2차원 평면그래프에 버블차트를 활용하여 사용자가 게재한 학술지를 표현한다. 이때 사용되는 사용자의 연구 정보를 바탕으로 연구분야 정보를 제공하고, 카테고리 유사도 측정 방법을 통하여 적절한 학술지를 상위 5개에 한해 추천한다.

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

C. 논문 구성

1장은 본 논문의 서론으로써, 연구 내용에 대한 연구 배경 및 목적에 대해 설명한다.

2장은 관련 연구로써, 본 논문에 대한 이론적으로 필요한 정보와 기존 연구들에 관한 동향을 파악한다.

3장은 Clarivate Analytics에서 배포하는 Journal Citation Reports(JCR) 서비스 내 학술지 카테고리 및 학술지 목록을 추출하여 관계를 설정한다. 이때, 한 한술지가 두 개 이상의 카테고리에 포함되는지 여부를 측정하여 이에 대한 관계를 설정하고, Google Scholar에서 제공하는 연구자 개개인별 연구 실적을 표현하고 있는 사용자 프로필 페이지에 대한 HTML 소스 추출하여, 불용어 제거 및 학술지명 추출 방법에 대해 기술한다. 이를 통해 앞서 정의한 학술지 리스트를 활용하여 사용자가 제출한 논문이 게재된 학술지의 빈도수 및 평가 지표 정보를 저장하는 방법에 대해 기술한다. 이는 사용자의 연구 분야를 보여주기 위한 데이터로 활용되며, 학술지 간 거리 측정에 사용된다. 또한, 추후 카테고리 유사도 측정 방법에서 카테고리 간 학술지 추천 방법에 사용된다.

4장은 지식베이스를 기반으로 카테고리명을 활용한 유사도 측정 방법을 통해 앞서 추출한 카테고리 및 학술지 데이터를 바탕으로 EigenValue 방법을 이용한 카테고리 유사도 측정 방법에 대해 기술한다. EigenValue 방법을 사용해 도출된 고유값을 카테고리 간 거리값으로 책정하여, 최적합 학술지 4편을 추천하는데 사용되는 척도가 된다.

5장은 실제 연구자 데이터를 통해 도출된 정보를 바탕으로 본 논문에서 제안하는 카테고리 유사도 측정 방법을 적용한 최적합 학술지를 추천하는 방법을 기술한다. 여기서 추천 학술지는 기존에 가장 많이 게재된 학술지를 바탕으로 이와 거리가 가까운 학술지를 추천하는 방법으로 연구자에게 보다 많은 학술지를 알려준다.

마지막으로 7장은 본 논문에 대한 결론 및 차후 연구에 대해 기술한다.

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

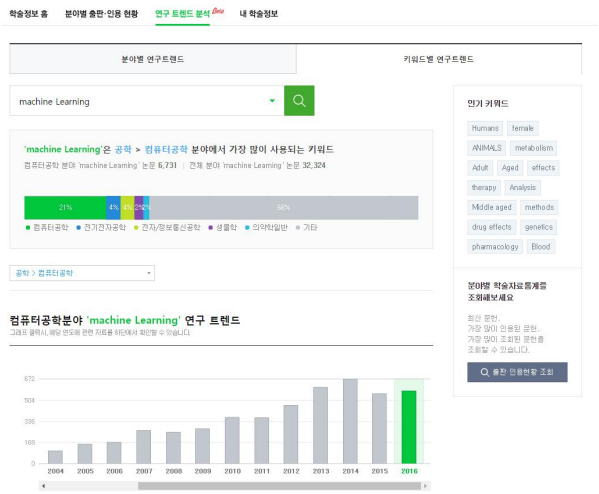
Ⅱ. 관련연구

A. 학술정보서비스 소개 및 현황

국내외 학술정보서비스는 포털사이트 형식으로 지속적인 기능 업데이트 등을 통해 다양한 관점에서 서비스를 제공하고 있다. 본 장에서는 국내외 학술정보검색서비스를 제공하고 있는 주요 사이트에 대해 알아보고자 한다.

1. 네이버 학술정보

네이버는 2017년에 기존 학술정보 서비스에서 '네이버 학술정보'로 개편하면서, 전문문서를 기준으로 한 학술자료를 중심으로 데이터베이스를 확대하였다. 이는 국내·외 논문 약 2억 1천만건, 참고문헌 약 14억건을 수집하였으며, 방대한 학술 데이터베이스를 바탕으로, 국내 연구자들이 필요로 하는 다양한 정보와 기능을 추가했다. 그리고, 검색 기능을 개선을 통한 주요 전문문서 검색속도를 향상 시켰으며, 대량의 참고문헌을 데이터베이스화하면서, 인용 수를 기준으로 한 논문 정렬방식을 추가하였다. 이와 같은 대량의 데이터베이스를 바탕으로 현재 베타테스트 중인 연구 트렌드 분석을 통해 연도별 연구 키워드 목록을 추출하여 각 분야별 주요 키워드 및 논문 수를 도식화하여 제공하고 있다.



[그림 2-1] 연도별 "Machine Learning" 대한 연구 트렌드 검색 결과

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

예를 들어, [그림 2-1]과 같이 각 연도별 컴퓨터 공학 분야 내 'Machine Learning' 키워드를 검색 시 해당 키워드와 관련된 논문 및 관련 분야에 대한 정보를 표현해준다. 하지만, 지속적으로 발전하고 있는 네이버 학술정보 서비스는 대부분 국내에서 사용되는 한계점이 존재하며, 타 학술정보서비스와 비교 시 추가적인 데이터 수집 및 개발이 필요하다.

2. 구글스칼라(Google Scholar)

구글스칼라의 세계적인 대표 검색엔진인 구글에서 서비스하는 학술 정보 서비스이다. 구글 스칼라는 기존 타 학술 정보 서비스와 비교하기 힘들 정도로 많은 자료를 보유하고 있다. 이는 저작권법에 따라 전문 문서 내용 중 일부를 발췌하여 내용을 제공하거나, 초록, 서지정보를 제공한다. 검색 자료는 논문, 서적 등 다양한 기준을 정의하여 검색 결과로 표현하고 있다.

특히, 구글스칼라는 출판사, 대학, 도서관, 연구소 등 학술연구 정보검색을 위한 공동작업을 진행함으로써 광범위한 데이터 공유협약을 체결하였다. 이를 통해 다양한 형태의 학술정보 및 Open Access를 통한 접근성을 접목하여 서비스 데이터 기반을 마련하였다[3].

구글의 메인화면은 검색란을 표시할 뿐 타 기능이 없는 것처럼 보이지만, 특정 검색어 또는 문장단위로 검색할 경우 각각의 형태별로 추가 검색이 가능하게 만들어져 있다. 예를 들면, 타 검색사이트와 마찬가지로 통합, 이미지, 뉴스, 동영상 등에 대한 분류를 기본으로 하고 있으며, 언어별 검색, 고급 검색, 날짜, 텍스트를 기준으로 한 완전일치 등과 같이 매우 세부적인 검색 결과를 받아볼 수 있도록 제작되었다. 더욱 놀라운 것은 검색란에 Filetype을 지정하면 해당 확장자를 가진 파일형태의 자료 유형만 검색 결과로 받아볼 수 있다.

또한 구글은 구글색인 서비스를 통하여 각 검색결과와 관련성이 높은 자료를 추가 링크 정보를 통해 나타내며, 누구나 관련자료를 쉽게 확인할 수 있다[4]. 구글이 전세계적으로 각광을 받는 가장 큰 이유는 페이지 랭크 알고리즘을 사용하여 검색된 문헌에 대한 랭크를 부여하였다. 이는 웹 페이지 내 존재하는 링크 정보를 기준으로 타 웹

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

페이지에서 링크되어 들어오는 횟수, 해당 웹 페이지에서 링크한 페이지들에 대한 정보를 바탕으로 구현된 알고리즘으로써 구글의 검색결과에 큰 영향을 주었다[5, 6]. 이를 바탕으로 응용연구를 통해 다양한 분야에서 사용되고 있다. 이처럼 구글 스칼라는 대표적 검색 포털사이트이자 학술정보검색 서비스를 제공한다. 따라서, 본 논문에서는 구글 스칼라 내 사용자 프로필 정보 중 연구 실적과 관련한 내용을 수집분석 하고, 이를 통해 사용자의 연구 분야와 관련된 학술지를 추천해주기 위한 기본 데이터로 활용하였다.

3. Microsoft Academic Search

마이크로소프트 학술검색은 베타버전으로 2006년 Live Academic Search에서 시작되었다. 현재 보유중인 색인 문서는 약 7백만 건을 넘었으며, 15개의 도메인으로 분류하여 자료를 제공하고 있다. 제공된 자료는 각 도메인별로 그 차이는 있으나, 타 학술정보서비스와 큰 차이는 발생하지 않는다[7].

앞서 설명한 구글 학술검색은 대부분의 분야를 포함하는 반면, MS 학술검색은 정보기술분야와 컴퓨터과학을 주 대상으로 한다. 하지만, 사용자에게 제공되는 서비스 부분에서는 논문의 인용정보, 연구자 간 네트워크 정보 등을 기본적으로 제공하며, 논문별 참고문헌에 대한 링크를 제공한다. 또한, 연구자 연결망과 검색결과를 좁힐 수 있는 필터링기능과 함께 분야별로 상위에 순위된 논문을 볼 수 있도록 하는 등 구글 스칼라보다 더 다양한 기능을 제공한다.

4. SCOPUS

Scopus는 Elsevier사에서 제공하는 세계 최대 규모의 학술 데이터베이스이다. 이는 공학, 과학, 의학, 기술, 예술・인문학 등의 분야를 포함하고 있으며, 약 19,500개의 저널에 대한 정보를 저장하고 있다. Scopus는 1996년부터 저널에 대한 인용정보를 제공하고 있으며, 전체 학술데이터 정보는 1965년부터 제공한다. 또한, Scopus는 Web of Science와 달리 전체 학술문헌의 50% 이상이 아시아 태평양 지역을 비롯한 유럽, 라틴아메리카 지역의 출판물로 구성됨으로써 전 세계 학술 데이터베이스의 2위를 차지 하고 있다. 대표적인 서비스로 사용자의 연구결과에 영향을 주는 논문별 인용정보를 실시간 제공하며, 각 저자별 H-Index를 포함한 저자 프로필 생성이 가능하다. 또한, 기관의 연구논문을 그룹핑하여 Affiliation Identifier 생성하여 정보를 제공한다[8].

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

이를 통해 논문에 대한 전반적인 분류가 가능하며, 사용자가 쉽게 확인할 수 있도록 정보를 제공한다. 현재 Scopus에서 보유중인 학술정보는 약 360만건의 학술발표 논문, 350여 편의 도서, 17,000여 건 이상의 Peer-Review 정보를 제공한다[9]. 하지만, 본 논문에서는 Scopus와 SCI등급과 차등관계로 인해 일반적으로 많은 학술지 정보를 제공받는 곳인 Web of Science 데이터를 사용한다.

B. 학술지 평가에 사용되는 지표

1. Impact Factor

Impact factor는 1955년 Eugene Garfield의 아이디어를 바탕으로 제안되었다[10]. 이를 통해, SCI(Science Citation Index, 과학인용색인)가 만들어졌으며, 학술지들에 대한 정량적 평가지표로 사용되기 시작했다. 이 지표는 매년 Thomson Reuter사의 Clarivte Analytics의 JCR(Journal Citation Reports)에 발표되어 연구자의 연구 영향력의 지표로 사용되고 있다. Impact Factor의 산출 방법은 해당 연도에 어떤 학술지의 그 전 2년 동안 해당 학술지에 출판되었던 논문들이 해당 연도에 받은 인용 횟수의 편당 평균값이다[11]. 구체적으로 다음과 같이 정의된다.

y년에 학술지 x의 impact factor는

$$Impact\,Factor = \frac{A}{B}$$

A = (y-1)년과 (y-2)년에 학술지 x에 출판된 모든 논문이 y년에 조사 대상 학술지들에 인용된 총 횟수

B = (y-1)년과 (y-2)년에 학술지 x에 출판된 모든 논문에 대한 총 인용수

여기서 인용된 학술지는 논문들만을 포함하며 기타 형식의 문서들은 포함하지 않는다. 반면에 A를 구할 때는 x에 출판된 모든 논문들에 대한 인용을 다 포함한다.

2. Eigenfactor Score

워싱턴대학교 생물학과 교수인 Carl T. Bergstrom에 의해 개발되었으며, The Metrics Eigen factor TM Project 수행과정에서 The Eigenfactor TM Algorithm을 개발하였다. Eigen Factor 웹사이트(eigenfactor.org)에서 1995년부터 2011년까지의 분석

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

자료를 제공하고 있으며, 2007년부터는 JCR을 통해 제공되고 있다[12].

이는 과학 커뮤니티에 대한 중요도를 측정한 점수로서 특정 학술지가 다른 학술지에 대하여 어느 정도의 영향력이 있는지를 평가하며, 인용 수에 기반하여 분석된다. 이는 Google에서 사용하는 PageRank 알고리즘과 유사하게, 학술지들을 랭크하는 방법이다.

Eigen Factor 웹사이트(eigenfactor.org)에서 115,000개의 reference를 모두 분석하여 제공하고, Science와 Social Science 및 다른 publications간의 연결관계도 고려된다. Eigenfactor Score는 학술지의 인용지수에 따라 점수의 가중치를 달리하여 측정한다. 다른 학술지에 미치는 영향력이나 인용한 학술지의 영향력에 따라 가중치가 달라지므로 Eigenfactor Score가 높을수록 학술지의 영향력이 크다고 할 수 있다.

C. WordNet

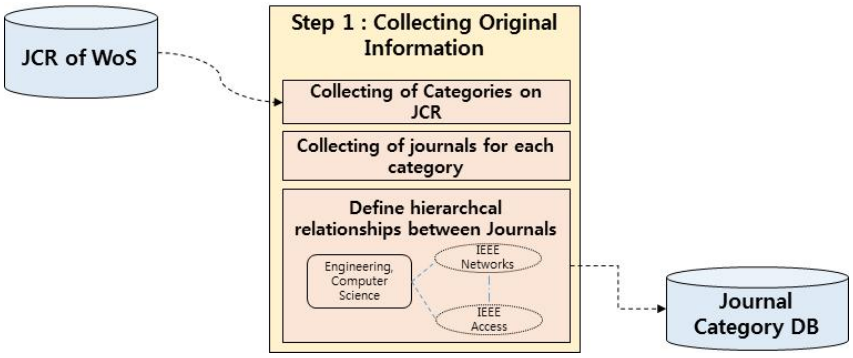
WordNet은 미국 프린스턴 대학 내 심리학 교수 A. 밀러의 인지 과학 연구소에서 만들어졌으며, 1985년부터 개발하기 시작하여 2006년도에 서비스를 시작한 전세계 대표적 의미 어휘목록이다. 이는 명사, 동사, 형용사, 부사를 구분하고 있으며, 각 단어별 고유 번호를 정의하고 있는데, 이를 동의어 집합(“Synset”)이라 부른다. 이를 기반으로 다의어, 이음동의어 정보를 함께 보여주고, 단어에 대한 간략한 설명과 문장을 표현해준다. 또한, 명사들 간의 개념정보를 표현하기 위해 상/하위 관계, 인스턴스, 도메인 등 관계를 분류하여 각 관계별 기호를 정의하였다[13, 14]. 그러므로 각 개념들 사이의 의미적 네트워크 관계를 구축할 수 있다. 본 논문에서는 상하위 관계 정보를 바탕으로 최상위 노드를 기준으로 각 단어별 거리 측정 방법을 통하여 카테고리명 및 학술지 내 Scope 정보를 기반으로 단어별 거리 측정 방법에 사용한다.

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

Ⅲ. 학술지 카테고리 분류 및 연구 실적 데이터분석 방법

본 장에서는 전 세계적으로 가장 사용빈도가 높은 Web of Science 시스템 내 JCR에서 제공하는 학술지 카테고리를 기반으로 전체 카테고리 구성 방법에 대해 기술하고, 이를 기반으로 각 학술지 간의 관계를 정의한다. 또한, 학술지 간의 거리를 측정함으로써 해당 학술지와의 유사성을 판별하는 방법을 제안한다. 본 장의 전체 구성도는 [그림 3-1]과 같다.



[그림 3-1] 데이터 전처리 및 Keyword Weight 산출을 위한 전체 구성도

A. 학술지 카테고리 분류 방법

1. JCR 저널 카테고리 수집 및 분석 방법

본 절에서는 현재 JCR 내에서 사용되는 저널 카테고리를 수집하여 해당 카테고리 내 저널간의 관계를 수집 및 분석하고자 한다. 2018년도를 기준으로 정의된 카테고리는 235개이며, 이는 [표 3-1]과 같다. 이는 전체 분야에 대한 저널지를 각각의 카테고리를 선정하여 분류하였으며, 이를 *'Subject Category'*라고 일컫는다. 예를 들어, 컴퓨터 과학 분야의 경우 *'Computer Science, Artificial Intelligence'*, *'Computer Science, Cybernetics'*, *Computer Science, Hardware & Architecture'*, *'Computer Science, Information Systems'*, *'Computer Science, Interdisciplinary*

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

Applications', '*Computer Science, Software Engineering*' 등 총 8개의 분야로 분류되어 있다. 이는 컴퓨터 과학분야 내에서도 다양한 연구 분야가 존재하기 때문에 세분화하여 카테고리를 선정해 놓은 것이다. [표 3-1]에서 '*Edition*'은 저널지의 SSCI, SCIE등급을 분류한 것이고, '*Journals*'는 해당 카테고리에 포함되어 있는 저널지의 수를 나타낸다. '*Total Cites*'는 해당 카테고리 내 포함된 저널지에 게재된 논문에서 발생한 전체 인용수를 나타내고, '*Median Impact Factor*'는 카테고리 내 전체 저널지의 Impact Factor를 기준으로 정렬하였을 때, 가장 중간이 되는 저널지의 Impact Factor 값을 나타내며, '*Aggregate Impact Factor*'는 기준 연도의 (year-1)과 (year-2)연도의 카테고리 내에 속하는 전체 학술지에 게재된 논문이 기준 연도에 인용된 총 인용횟수를 (year-1)과 (year-2)연도에 같은 학술지들에 게재된 모든 인용 논문으로 나눈 것이다.

[표 3-1] JCR에 정의된 저널 카테고리

Rank	Category	Edition	Journals	Total Cites	Median Impact Factor	Aggregate Impact Factor
1	ECONOMICS	SSCI	353	905,730	1.112	1.766
2	MATHEMATICS	SCIE	310	494,556	0.704	0.855
3	BIOCHEMISTRY & MOLECULAR BIOLOGY	SCIE	293	3,625,819	2.906	4.281
4	MATERIALS SCIENCE, MULTIDISCIPLINARY	SCIE	285	3,451,318	2.008	4.641
5	NEUROSCIENCES	SCIE	261	2,346,383	3.047	4.015
⋮	⋮	⋮	⋮	⋮	⋮	⋮
232	MICROSCOPY	SCIE	10	33,505	1.659	1.809
233	ANDROLOGY	SCIE	6	8,410	1.785	2.234
233	GREEN & SUSTAINABLE SCIENCE & TECHNOLOGY	SSCI	6	17,898	1.984	2.124
235	TRANSPORTATION	SCIE	1	3,220	0.983	0.983

이처럼 많은 카테고리 내 존재하는 저널들은 그 수가 매우 많다. 예를 들어, '*Computer Science, Information Systems*'에 해당하는 저널의 수는 148개로 구성되어 있으며, 해당 카테고리가 생성된 후 현재까지의 정보들이 정리되어 있다. 이는 [표 3-2]와 같다.

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

[표 3-2] *'Computer Science, Information Systems'* 카테고리 정보

Year	Edition	Journals	Articles	Total Cites	Median Impact Factor	Aggregate Impact Factor	Aggregate Immediacy Index	Aggregate Cited Half-Life	Aggregate Citing Half-Life
2017	SCIE	132	14,561	497,284	1.725	3.079	0.807	7.0	7.7
2016	SCIE	133	13,960	457,382	1.807	2.997	0.587	7.2	7.9
2015	SCIE	130	12,773	331,014	1.403	2.250	0.424	7.2	8.0
2014	SCIE	123	11,450	301,413	1.406	2.134	0.388	7.3	8.1
2013	SCIE	121	10,620	259,711	1.167	2.003	0.320	7.5	8.1

[표 3-3]은 *'Computer Science, Information Systems'* 카테고리 내 학술지의 목록과 해당 정보를 나타낸다. 이는 Impact Factor가 높은 순으로 정렬되어 있으며, 각각의 학술지별 인용 수 및 EigenFactor Score를 표현하고 있다.

[표 3-3] *'Computer Science, Information Systems'* 내 포함된 학술지 정보

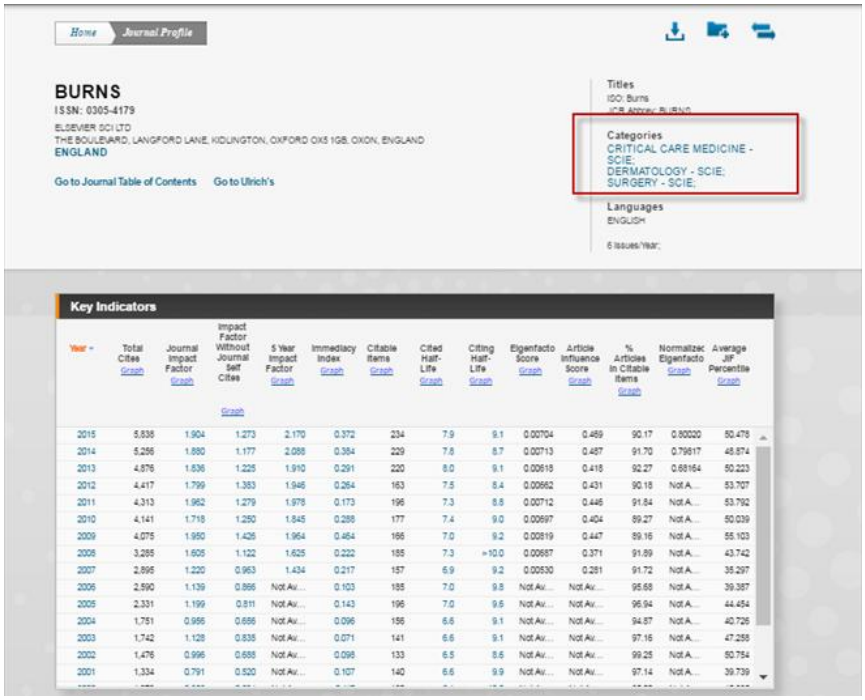
Rank	Full Journal Title	Total Cites	Journal Impact Factor	Eigenfactor Score
1	INTERNATIONAL JOURNAL OF COMPUTER VISION	16,572	11.541	0.025000
2	IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE	46,505	9.455	0.064000
3	IEEE Transactions on Cybernetics	9,424	8.803	0.033000
4	IEEE TRANSACTIONS ON FUZZY SYSTEMS	12,946	8.415	0.021000
5	IEEE TRANSACTIONS ON EVOLUTIONARY COMPUTATION	10,682	8.124	0.009000
⋮	⋮	⋮	⋮	⋮
129	COMPUTING AND INFORMATICS	367	0.410	0.000000
130	INTERNATIONAL JOURNAL OF SOFTWARE ENGINEERING AND KNOWLEDGE ENGINEERING	360	0.397	0.000000
131	ACM Transactions on Asian and Low-Resource Language Information Processing	23	0.339	0.000000
132	Traitement du Signal	51	0.057	0.000000

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

2. 카테고리 내 학술지 간 관계 정의

앞서 기술한 Web of Science에서 제공하는 전체 카테고리에 대한 저널의 관계 정의는 해당 사이트에서 분류해놓은 카테고리를 기반으로 하며, 하위에 속해 있는 저널을 기반으로 타 카테고리에 속해있는지 여부를 판별하여 관계를 정의한다. 예를 들면, [그림 3-2]와 같다. [그림 3-2]에서는 각각의 카테고리 'A', 'B', 'C'가 있다고 가정할 때 해당 카테고리 내 포함된 학술지의 수는 4개, 4개, 1개이다. 이때 카테고리 'A', 'B', 'C' 내에 포함되어 있는 각각의 학술지는 카테고리의 하위 개념을 갖게 되며, 각각의 카테고리 내 포함된 학술지 간 관계가 존재하는 경우 해당 카테고리 간 연관성이 있는 것으로 간주하였다. 이는 한 학술지가 1개 이상의 카테고리를 가지고 있는 경우가 있고, 해당 학술지에서 세부적인 분야의 논문만 취급하는 것이 아니라는 것을 확인하여 이와 같이 관계를 형성한 것이다. [그림 3-2]는 JCR에서 *'Burns'* 학술지의 정보를 표현하고 있는 모습을 보여준다. 이는 Elsevier에서 출판한 학술지로 2.134의 Impact Factor로 책정된 일반적인 학술지이다. 특히, 붉은색으로 표현된 곳은 해당 학술지가 소속된 카테고리 정보를 보여주고 있다. 이를 통해 각각의 학술지의 카테고리 정보를 추출할 수 있고, 해당 카테고리를 기준으로 학술지 간 계층관계를 형성할 수 있다.

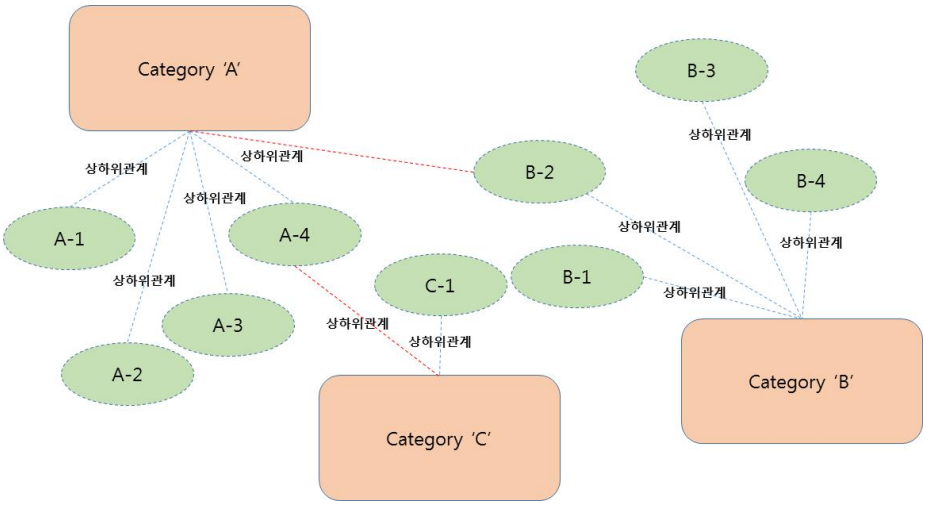


[그림 3-2] JCR 내 'BURNS' 학술지의 카테고리 정보

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

예를 들어, [그림 3-3]과 같이 'A-4'라는 명칭을 가진 학술지의 경우 카테고리 'A'와 카테고리 'C'에 포함되어 있는 것을 알 수 있으며, 'B-2'라는 명칭을 가진 학술지의 경우 카테고리 'B'와 카테고리 'A'에 공통적으로 포함되어 있다는 사실을 확인 하였으며, 이를 Python 도구를 활용하여 전체 235개의 카테고리와 이에 포함된 학술지들 간에 계층 관계 및 다중 포함관계를 분석 및 정의 한다.



[그림 3-3] 각 카테고리별 학술지 계층 구조 및 학술지 간 관계 형성 예제

위 과정을 통해 얻어진 결과는 [표 3-4]와 같다. 이는 *'Computer Science'* 분야의 카테고리 내에 포함되어 있는 학술지가 2개의 카테고리에 동시 포함되는 개수를 나타내고 있다. 즉, *'Computer Science, Artificial Intelligent'* 분야의 학술지는 각각 *'Computer Science, Hardware & Architecture'* 분야의 학술지와 공통적으로 포함된 학술지의 개수는 7개가 있는 것으로 확인된 것이다.

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

[표 3-4] 두 개의 카테고리 사이에 동시에 포함되는 저널의 수

	CSAI	CSCY	CSHA	CSIS	CSIA	CSSE	CSTM
CSAI (132)	-	7	1	17	10	6	21
CSCY (22)	-	-	0	1	0	1	2
CSHA (52)	-	-	-	13	4	18	16
CSIS (148)	-	-	-	-	8	30	19
CSIA (105)	-	-	-	-	-	6	5
CSSE (104)	-	-	-	-	-	-	5
CSTM (103)	-	-	-	-	-	-	-

CSAI : Computer Science, Artificial Intelligent
CSCY : Computer Science, Cybernetics
CSHA : COMPUTER SCIENCE, HARDWARE & ARCHITECTURE
CSIS : COMPUTER SCIENCE, INFORMATION SYSTEMS
CSIA : COMPUTER SCIENCE, INTERDISCIPLINARY APPLICATIONS
CSSE : COMPUTER SCIENCE, SOFTWARE ENGINEERING
CSTM : COMPUTER SCIENCE, THEORY & METHODS

B. Google Scholar를 이용한 개인 연구 실적 수집 방법

1. HTML 문서 전처리 및 정보 추출 방법

본 절에서는 비정형 데이터 전처리 과정에 대해 설명한다. 본 논문에서 사용하고자 하는 문서는 Google Scholar에서 제공하는 개인 연구실적 및 각각의 개인 평가지표를 확인할 수 있도록 제공하고 있는 웹 문서를 수집한다. 그 후 HTML Parser를 사용하여 불필요 요소를 제거한다. 여기서 불필요 요소는 웹 문서를 작성할 때 기본적으로 사용되는 HTML 태그, Java Script, CSS 등 웹 문서의 기반이 되는 요소들을 제거한다.

현재까지 개발된 HTML Parser는 매우 많이 존재한다. 하지만, 본 연구에서 사용하는 Python 도구를 사용하여 HTML Parser를 이용하여 문서 내 개인의 연구 실적 중 저널에 게재한 논문만 추출하고자 한다. 웹 문서 내 텍스트 추출 단계를 설명하기 위해 특정 저자의 Google Scholar 정보를 기준으로 텍스트 추출 방법에 대해 예제를 제시한다. 이는 [표 3-5]와 같다.

Collection @ chosun

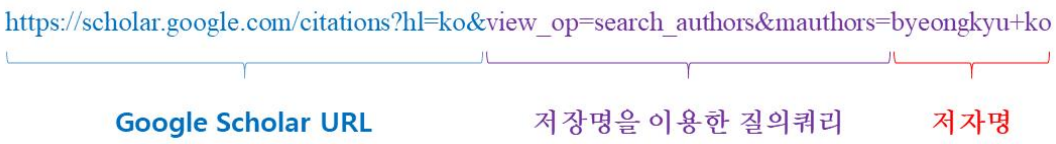
조선대학교
CHOSUN UNIVERSITY

[표 3-5] Google Scholar 내 개인연구실적 정보 추출 단계

단계 1. Google Scholar 내 개인 연구 실적정보 페이지 Link 입력 단계 2. Python HTML Parser를 이용하여 특정 태그 내 정보추출 단계 3. 추출 된 정보 내 학술지명을 기준으로 데이터 분류 단계 4. 추출된 텍스트는 txt파일 형식으로 분리 저장

a. Google Scholar 내 개인 연구 실적정보 페이지 Link 입력

저자의 성명을 기준으로 Google Scholar 프로필에 관한 소스 정보를 가져오기 위해 웹에서 제공되는 URL값이 필요하다. 이 때문에 Google Scholar에서 사용하는 질의 쿼리가 필요하며, 해당 쿼리문은 [그림 3-4]와 같다. 이를 이용하여 Google Scholar 프로필 정보를 수집하는데 사용한다.



[그림 3-4] Google Scholar에서 사용하는 프로필 검색 쿼리문

이 과정을 거쳐, 최종적으로 추출해 낸 해당 저자의 프로필은 한명, 또는 다수의 이용자가 나타날 수 있다. 이를 고려하여 본 연구에서는 한명의 저자가 검색결과로 도출된다는 가정 하에 연구를 진행한다.

위 쿼리문으로 검색한 후 한명의 검색결과가 도출되었을 시 추가적인 URL이 필요하다. 이 이유는 Google Scholar에서 제공하는 프로필 페이지로 이동할 때 해당 사용자의 고유번호로 이동을 하기 때문이다. 이를 통해 최종적으로 추출되는 URL값은 [그림 3-5]와 같다.

Collection @ chosun

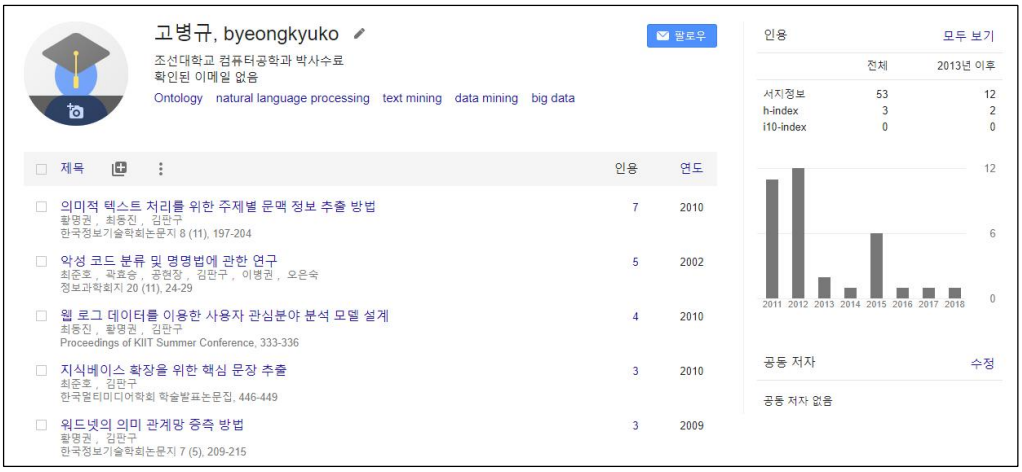
조선대학교
CHOSUN UNIVERSITY

[그림 3-5] HTML Parser에서 사용할 최종 URL

b. Python HTML Parser를 이용하여 특정 태그 내 정보추출

Google Scholar에서 제공하는 연구자 프로필정보를 통해 연구자의 연구실적 및 전체 피인용 수, h-index, i10-index 정보를 확인할 수 있으며, 추가적으로 최근 8년 내 전체 피인용 수를 확인할 수 있도록 만들어 졌다.

<그림 3-6>과 같이 개인의 연구정보를 확인할 수 있도록 필요정보들을 도출하여 표현해준다. 또한, 개인의 논문정보를 수집하여 자동으로 출력해준다.



[그림 3-6] Google Scholar에서 제공하는 개인 프로필 정보

본 연구에서는 웹 문서에서 표현하고 있는 전체 정보 중 해당 저자의 논문 리스트를 추출하고자 한다. 이를 위해 해당 페이지의 HTML 소스 정보를 확인하여 추출하고자 하는 정보가 어떤 태그를 통해 정보를 표현하는지 분석이 필요하다. 이를 위해 <그림 3-7>은 Google Scholar의 웹 문서 페이지의 HTML 소스를 보여준다.

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

[그림 3-7] 개인 프로필 정보 웹 문서의 HTML 소스 정보

앞서 말했듯이 기본적인 HTML 문서의 구조는 Header, Title, Style, Body 정보로 이루어져 있으며, 해당 연구에서 필요한 정보를 추출하고자 하는 태그는 Body 태그 내에 다양한 속성 및 클래스로 표현되어 있다. [그림 3-8]은 HTML 소스 내 필요한 정보가 포함된 태그를 분석한 결과를 나타낸다.

[그림 3-8] 개인 프로필 정보 웹 문서의 HTML 소스 정보

[그림 3-8]과 같이 HTML 소스정보와 웹 문서 상 표현된 정보를 분석한 결과 각각의 표시된 번호끼리 연관되어 있는 것을 확인할 수 있다. 즉, 본 연구에서 사용하고자 하는 데이터는 ①번 데이터 내에서 해당 프로필의 저자명, 저자소속, 저자 이메일 정보를 추출한다. 그 이유는 첫째, 저자명의 경우 해당 저자에게 정보를 알려주기 위해 기

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

본적으로 필요한 것이며, 둘째로 저자소속은 동명이인의 경우 소속정보 및 이메일을 통해 명확하게 저자를 파악하기 위해 사용되기 때문이다. ②번, ③번은 해당 프로필의 저자와 관련한 모든 논문 리스트이며, 이는 본 연구에서 가장 중요하게 사용될 데이터이기 때문에 각각을 분류하여 수집하기 위한HTML Parser 규칙이 필요하다. [표 3-6]은 각각의 불필요 데이터를 제거함과 동시에 중요 데이터 추출을 위한 HTML Parser 규칙을 정의한 것이다.

[표 3-6] HTML Parser를 사용하기 위한 추출 규칙 정의

규칙 1	‘div’ 태그 내 ‘id’속성값을 ‘gsc_prf_in’으로 설정이 되어 있으며, 이는 해당 프로필의 저자명을 나타낸다.
HTML Parser	'div', id='gsc_prf_in'
규칙 2	‘div’ 태그 내 클래스명이 ‘gsc_prf_il’로 설정이 되어 있으며, 해당 저자의 소속정보를 나타낸다.
HTML Parser	'div', class_='gsc_prf_il'
규칙 3	‘div’ 태그 내 ‘id’속성값이 ‘gsc_prf_ivh’로 설정이 되어 있으며, 해당 저자의 이메일 정보를 나타낸다.
HTML Parser	'div', id='gsc_prf_ivh'
규칙 4	‘a’ 태그 내 클래스명이 ‘gsc_a_at’로 설정이 되어 있기 때문에 다음과 같이 Html Parser 규칙을 설정한다.
HTML Parser	'a', class_='gsc_a_at'
규칙 5	‘div’ 태그 내 클래스명이 ‘gs_gray’로 설정이 되어 있기 때문에 다음과 같이 HTML Parser 규칙을 설정한다.
HTML Parser	'div', class_='gs_gray'

위 규칙들 중 동일한 ‘div’ 태그를 사용할 시 ‘id’ 속성값을 기준으로 데이터를 추출하는 것과 클래스명을 기준으로 데이터를 추출하는 방법으로 설정되어 있다. 이는 ‘id’ 값과 클래스명은 중복으로 사용할 수 있기 때문에 데이터를 수집하는데 있어, 불필요 데이터를 추출할 가능성이 있다. 이를 방지하기 위해 같은 클래스명을 가진 데이터를 추출했을 시 (규칙 2)와 (규칙 3)은 추가적으로 ‘id’ 속성값을 통해 각각의 이메일 정보 및 소속정보를 추출한다. 또한, ‘a’ 태그와 ‘div’ 태그 내 클래스명도 중복으로 사용되는 경우가 있지만, 해당 웹 문서에서는 각각 다르게 표현되어 있다.

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

하지만, 위 2개의 태그 뿐만 아니라 다른 태그에서도 같은 클래스명을 사용할 경우가 있고 해당 클래스 내 속성값을 상속하기 위해 많이 사용되는데, 이 같은 경우 데이터 추출에 있어 타 정보가 함께 흘러 들어올 수 있다. 따라서, HTML Parser를 통해 중복 데이터 또는 추가 불필요 요소가 추출되는지 확인 및 검토를 거친후 최종적으로 규칙을 설정한 것이다.

c. 추출 된 정보 내 학술지명을 기준으로 데이터 분류 및 데이터 저장

위 (2)의 과정에서 정의한 규칙을 기반으로 데이터를 추출하며, 그 과정은 [표 4-3]과 다음과 같다.

[표 3-7] Google Scholar 프로필 내 주요 데이터 추출 과정

순서	과정	처리 결과
1	Google Scholar 프로필 정보 수집	Detecting Web based DDoS Attack using MapReduce operations in Cloud Computing Environment. J Choi, C Choi, B Ko, D Choi, P Kim J. Internet Serv. Inf. Secur. 3 (3/4), 28-37 55 2013 ~~Research for the pattern analysis of individual interest using SNS Data: focusing on Facebook~~ ~~B Ko, D Choi, J Kim, E Lee, C Choi, J Hong, P Kim~~ ~~2014 Eighth International Conference on Innovative Mobile and Internet … 2 2014~~ ⋮
2	논문종류 구분	Detecting Web based DDoS Attack using MapReduce operations in Cloud Computing Environment. J Choi, C Choi, B Ko, D Choi, P Kim J. Internet Serv. Inf. Secur. 3 (3/4), 28-37 55 2013 A method of DDoS attack detection using HTTP packet pattern and rule engine in cloud computing environment J Choi, C Choi, B Ko, P Kim Soft Computing 18 (9), 1697-1703 43 2014
3	구성요소별 분류 및 데이터 병합	Title : A method of DDoS attack detection using HTTP packet pattern and rule engine in cloud computing environment Author(s) : J Choi, C Choi, B Ko, P Kim Journal : Soft Computing Impact Factor : 2.367 EigenFactor Score : 0.008000 Related Category : Computer Science, Aritificial Intelligent, Computer Science, Interdisciplinary Applications

[표 3-7]에서 전체 과정을 설명하면, (1) Google Scholar 프로필 정보 수집 시 해당 사용자가 지금까지 게재 또는 발표한 논문, 보고서, 학위논문 등의 데이터를 수집할 수 있다. 본 연구에서 이용하는 카테고리 정보는 국제 학술지가 기본이 되기 때문에 학위 논문, 보고서, 발표논문의 경우는 배제가 되어야 한다. 따라서, 기존에 JCR 내에

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

서 추출한 논문 리스트를 활용하여 리스트 내 존재하지 않거나, ‘Conference’, ‘Thesis’, ‘Report’ 등의 불용어 리스트를 만들어 해당 단어가 존재하는 경우는 모두 삭제 처리 하였다. 이를 통해 (2) 논문종류 구분에서 최종적으로 국제 학술지와 관련한 논문 리스트를 수집할 수 있다. 그리고 프로필 사용자가 속해 있는 논문들은 1저자 또는 공저자를 구분하지 않고 전체 논문에 대한 학술지 목록을 추출하며, 최종적으로 구성요소 분류 및 데이터 병합 시 해당 논문과 관련한 추가 정보들을 병합하여 .txt 파일 형식으로 저장한다. 이때, 저장하는 파일명은 Google Scholar 프로필의 사용자명으로 저장한다.

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

2. 연구 실적 분석결과 표현 방법

[표 3-8]은 앞서 기술한 특정 연구자의 연구실적을 수집하여 그 결과를 도출하였으며, 각 학술지별 거리 측정에 사용하기 위해 Impact Factor값과 EigenFactor Score를 포함하여 정리한 결과를 나타낸다.

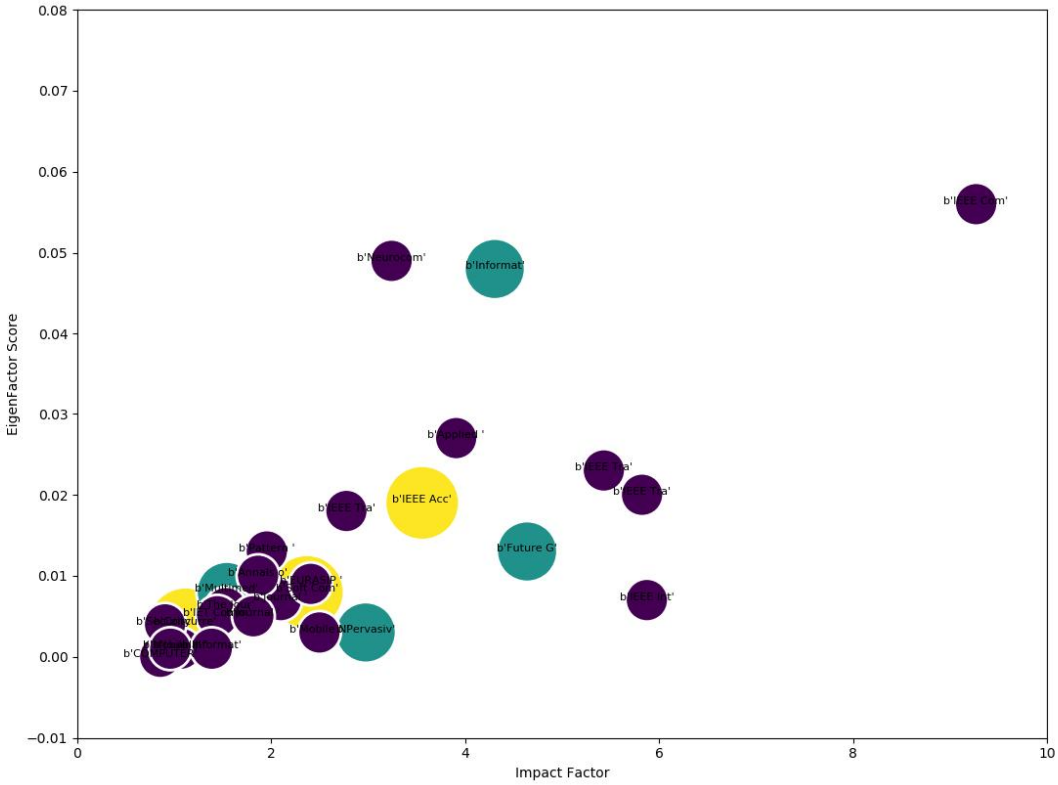
[표 3-8] 학술지별 빈도수 측정 결과 및 학술지 지표

Journal Title	Freq.	IF	ES	Journal Title	Freq.	IF	ES
Concurrency and Computation: Practice and Experience	3	1.114	0.004	IEEE Transactions on Industrial Informatics	1	5.43	0.023
Soft Computing	3	2.367	0.008	IEEE Transactions on Information Forensics and Security	1	5.824	0.02
IEEE Access	3	3.557	0.019	IET Communications	1	1.443	0.005
Future Generation Computer Systems	2	4.639	0.013	Applied Soft Computing	1	3.907	0.027
Pervasive and Mobile Computing	2	2.974	0.003	Mobile Networks and Applications	1	2.497	0.003
Multimedia Tools and Applications	2	1.541	0.008	IEEE Communications Magazine	1	9.27	0.056
Information Sciences	2	4.305	0.048	Security and Communication Networks	1	0.904	0.004
COMPUTER SYSTEMS SCIENCE AND ENGINEERING	1	0.855	0	Annals of Operations Research	1	1.864	0.01
Journal of Universal computer science	1	1.066	0.001	Journal of Parallel and Distributed Computing	1	1.815	0.005
Pattern Recognition Letters	1	1.954	0.013	IEEE Internet of Things Journal	1	5.874	0.007
The Journal of Supercomputing	1	1.532	0.006	Informatica	1	1.386	0.001
IEEE Transactions on Knowledge and Data Engineering	1	2.775	0.018	EURASIP Journal on Wireless Communications and Networking	1	2.407	0.009
Journal of medical systems	1	2.098	0.007	Neurocomputing	1	3.241	0.049

[표 3-8]에 기재된 데이터를 기준으로 2차원 평면좌표 내 [그림 3-9]와 같이 표현한다. 이때, 기준이 되는 요소는 일반적으로 학술지의 영향력 및 중요도를 나타내는 Impact Factor(X축)과 EigenFactor Score(Y축)으로 선정한다. 물론 이외에 피인용수 및 인용수와 같은 다양한 지표들도 존재하지만 Impact Factor와 EigenFactor Score를 계산할 때 모두 포함이 되어 있기 때문에 두 가지의 요소를 기준으로 표현한다.

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY



[그림 3-9] Google Scholar 내 연구 실적을 기반으로 한 학술지 분포도

[그림 3-9]는 [표 3-8]의 데이터를 기반으로 나타낸 사용자 연구실적에 대하여 좌표 평면상에 나타낸 결과이다. 각각의 원 크기는 학술지별 빈도수에 따라 크기가 변경되며, 빈도가 높을수록 원의 크기는 커진다. 색상은 빈도수를 기준으로 변경되며, 같은 빈도수를 가진 학술지의 색상은 동일한 색상으로 지정된다. 이와 같은 방법을 통해 사용자가 자신의 연구결과를 쉽게 이해할 수 있다.

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

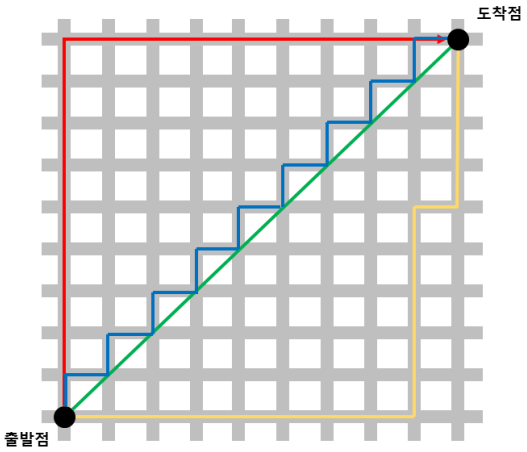
Ⅳ. 카테고리 유사도 측정 방법

본 논문의 목적인 학술지 추천을 위한 방법의 일환으로 연구자가 가장 많은 논문을 게재한 학술지를 기준으로 해당 학술지가 포함된 카테고리 내 전체 학술지에서 대한 거리 측정방법에 대해 기술한다. 그리고 학술지가 포함된 카테고리가 두 개 이상일 경우 다수의 카테고리는 연관이 있다고 판단하여 카테고리 간의 거리를 측정하기 위해 EigenValue를 기반으로 한 카테고리 유사도 측정 방법에 대해 기술한다.

A. 거리 측정 알고리즘 소개

1. 맨해튼 거리(Manhattan Distance)

맨해튼 거리 알고리즘은 19세기에 헤르만 미코프스키가 고안했으며, 유클리디안 거리 공식과 비교되는 공식으로 'L1 Distance'라는 이름으로 쓰이기도 한다. 해당 알고리즘의 맨해튼이라는 이름은 뉴욕 내 맨해튼의 전경을 항공기의 관점에서 볼 때, 'Street'과 'Avenue'가 사각형 격자로 잘 정렬된 모습으로부터 유래되었다. 이는 앞서 말했듯이 격자로 이루어진 지도상에서 출발점부터 도착점까지의 직선 거리가 아닌 최단거리를 구하는 공식이다. 맨해튼 거리 알고리즘을 그림으로 나타내면 [그림 4-1]과 같다[15, 16, 17, 20].



[그림 4-1] 맨해튼 거리 알고리즘 예

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

[그림 4-1]과 같이 출발점부터 도착점까지 이동할 수 있는 방법은 그림에서 예시한 부분보다 더 다양한 방법이 있다. 일반적으로 녹색선외에 가장 빠른 방법이라고 생각하는 것은 파란색선이라고 생각할 수 있다. 하지만, 맨해튼 거리 알고리즘을 이용하여 거리를 계산해 봤을 경우 위 4가지 방법 모두 같은 거리를 이동한다는 계산 결과가 나오게 된다. 즉, 위와 같이 같은 크기의 사각형 격자로 이루어진 짧은 선들로 이루어진 경우 해당 길이만큼 동일한 거리로 옮겨져 만들어지기 때문이라는 것이다. 이는 가장 간단한 방법으로 최단거리를 찾을 수 있으며 <수식 1>을 사용해 계산할 수 있다.

$$\begin{aligned} L_1 &= |x_1 - x_2| + |y_1 - y_2| \\ &= \sum_{i=1}^{n} |a_i - b_i| \end{aligned}$$ <수식 4-1>

<수식 4-1>은 기본적인 맨해튼 거리 알고리즘을 n차원으로 확장시켜 일반화 시킨 경우까지 도출한 결과를 나타낸다. 즉, 모든 점에 대해 각각의 x축 좌표값과 y축 좌표값을 빼준 값을 더한 결과를 나타낸다고 할 수 있다.

2. 유클리디안 거리(Euclidean Distance)

유클리디안 거리 알고리즘은 맨해튼 거리 알고리즘과 유사한 두 점사이의 최단 거리를 구하기 위해 고안된 알고리즘이며, L2 Distance라고도 불린다. 기존 피타고라스 정리를 사용하여 두점 사이의 거리를 구하는 방법을 차원 수를 적용하여 일반화 시킨 것과 유사하다. 유클리디안 거리 알고리즘을 수식으로 나타내면 <수식 4-2>와 같다[17].

$$\begin{aligned} L_2 &= \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_n - q_n)^2} \\ &= \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2} \end{aligned}$$ <수식 4-2>

<수식 4-2>와 같이 유클리디안 거리 알고리즘을 나타낼 수 있다. 하지만, 위 식으로 얻은 결과값은 거리의 최대값이 정해지지 않기 때문에 두 점사이의 유사도를 측정하는데 어려움이 있다. 따라서, 0 ~ 1사이의 값을 갖도록 정규화 하기 위해 1을 더하여 역수를 취해주는 방식을 취한다. <수식 4-3>은 이를 유클리디안 거리 정규화식을 나타낸다. 이를 통해 최종적으로 0 ~ 1사이의 값을 갖는 결과를 얻을 수 있으며, '1' 에 가

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

까울수록 두 객체사이의 유사도가 높고, '0'에 가까울수록 유사도가 낮다고 할 수 있다[18, 19, 20].

$$Nomalized\,L_2 = \frac{1}{1+\sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}}$$ <수식 4-3>

3. 민코프스키 거리(Minkowski Distance)

민코프스키 거리 알고리즘은 맨해튼 거리(L1 Distance)와 유클리디안 거리(L2 Distance)를 한번에 표현한 공식으로써, Lm Distance로 불린다. 민코프스키 거리 알고리즘을 수식으로 나타내면 <수식 4-4>와 같다. 민코프스키 공식에서 m의 값이 1인 경우 맨해튼 거리 알고리즘이 되며, m의 값이 2이면 유클리디안 공식이 된다[20].

$$L_m = \sqrt[m]{\sum_{i=1}^{n}(|a_i - b_i|)^m}$$ <수식 4-4>

이와 같이 다양한 거리 측정 알고리즘이 사용되고 있는데, 맨해튼 거리 알고리즘은 직선 거리가 아닌 특정 사물을 회피한다는 전제하에 최단거리를 측정하는 알고리즘이기 때문에 본 연구에 적용하기에는 다소 차이점이 존재한다. 또한, 민코프스키 알고리즘은 확률 분포를 고려하여 계산된다는 점으로 이도 마찬가지로 해당 알고리즘을 적용하는데는 어려움이 있다. 따라서, 본 연구에서 지금까지 가장 많이 사용되고 있는 유클리디안 거리 알고리즘을 사용하여 카테고리 내 학술지 간 거리를 측정하는 방법으로 사용하고 이를 통해 가장 근접해 있는 학술지를 저자에게 추천해 주는 방법을 사용한다.

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

B. 거리 측정 알고리즘을 사용한 학술지 간 거리 측정 방법

Google Scholar 내 연구 실적 분석을 통해 얻는 데이터를 바탕으로 2차원 좌표평면상 데이터 표현 방법과 본 연구에서 사용하고자 하는 거리 측정 방법에 대해 기술하였다. 이를 기반으로 한 개의 카테고리 내 포함된 전체 학술지와의 관계 추출을 위해 유클리디안 거리 측정 알고리즘을 사용하여 학술지 간의 거리를 측정한다. 여기서 사용하는 유클리디안 거리 측정 알고리즘은 최단거리를 직선으로 계산하여 도출하는 방법으로써, 앞서 설명한 바 있다. 예를 들어, "Soft Computing" 학술지를 기준으로 거리를 측정하였을 때, [표 4-1]과 같이 자기 자신과의 거리는 "1" 값을 갖고, 타 학술지와의 거리는 Impact Factor(x축)과 EigenFactor Score(y축) 데이터를 기반으로 두 점 사이의 거리를 도출할 수 있다. 이를 통해 도출된 결과는 [표 4-1]과 같다.

[표 4-1] 학술지별 빈도수 측정 결과 및 학술지 지표

학술지명	거리값	학술지명	거리값
Soft Computing	1	Neurocomputing	0.436108
Pattern Recognition Letters	0.954113	IEEE Communications Magazine	0.419639
Annals of Operations Research	0.826446	Journal of medical systems	0.37579
Journal of Universal computer science	0.794262	Multimedia Tools and Applications	0.34965
EURASIP Journal on Wireless Communications and Networking	0.786153	Mobile Information Systems	0.319747
Applied Soft Computing	0.752445	Future Generation Computer Systems	0.29044
IEEE Transactions on Knowledge and Data Engineering	0.705216	Mobile Networks and Applications	0.263637
Information Sciences	0.700762	COMPUTER SYSTEMS SCIENCE AND ENGINEERING	0.238589
IEEE Internet of Things Journal	0.587889	Pervasive and Mobile Computing	0.220994
Journal of Parallel and Distributed Computing	0.571421	IEEE Transactions on Information Forensics and Security	0.18811
The Journal of Supercomputing	0.543464	IET Communications	0.175131
IEEE Transactions on Industrial Informatics	0.504031	Informatica	0.173611
IEEE Access	0.443851	Security and Communication Networks	0.109216

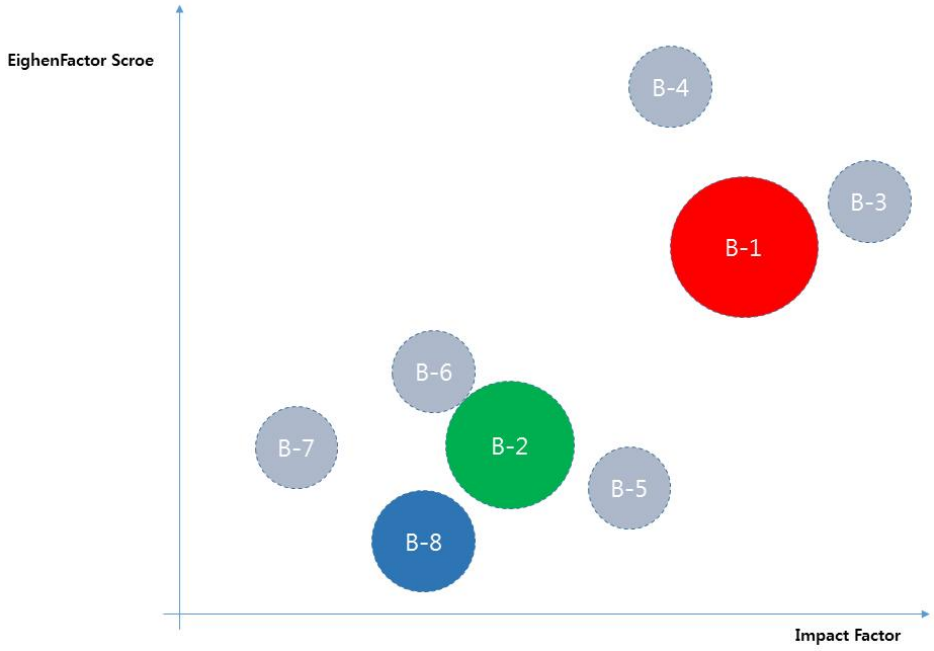
이는 특정 연구자가 게재한 학술지에 대한 지표를 각각의 점으로 표현하였으며, 이를 통해 각 학술지간 거리를 측정할 수 있다. [표 4-1]은 해당 연구자가 가장 많은 논문을 게재한 'Soft Computing' 학술지를 기준으로 가장 유사도가 높은 순으로 정렬한 결과를 나타낸 것이다.

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

해당 거리값을 기준으로 앞서 표현한 2차원 좌표평면상의 데이터와 비교해보면 다음과 같은 결과를 도출할 수 있다. 예를 들어, ‘Soft Computing’, ‘Concurrency and Computation: Practice and Experience’, ‘IEEE Access’는 각각 3편씩의 논문을 제출하였다. 그렇기 때문에 특정 한 개의 학술지를 기준으로 거리값을 측정하기에는 그 형평성이 맞지 않다. 또한, 학술지를 추천해주기 위해서는 추천하고자 하는 학술지가 해당 저자가 게재했던 학술분야와 공통점이 있어야 한다. 하지만 위의 [표 4-1]과 같이 다양한 연구분야(Medical, Security 등)의 논문을 제출한 연구자의 경우 이를 추천하기가 매우 까다롭기 때문에 해당 학술지가 가장 많이 포함된 연구분야를 기준으로 추천을 해야 하기 때문에 기존 연구실적물에 포함된 단어를 분석하여 각 단어별 상위 노드를 검색한 후 해당 카테고리와 연관된 단어가 많은 연구분야로 학술지를 추천한다.

학술지 추천을 위해 고려해야할 사항 중 해당 연구자에게 어느 범위까지 추천을 해야 하는지에 관한 경우이다. 예를 들면 [그림 4-2]와 [그림 4-3]과 같은 경우 추천을 하기 위한 기준을 정해야 한다.



[그림 4-2] Google Scholar 내 연구 실적을 기반으로 한 저널 분포도

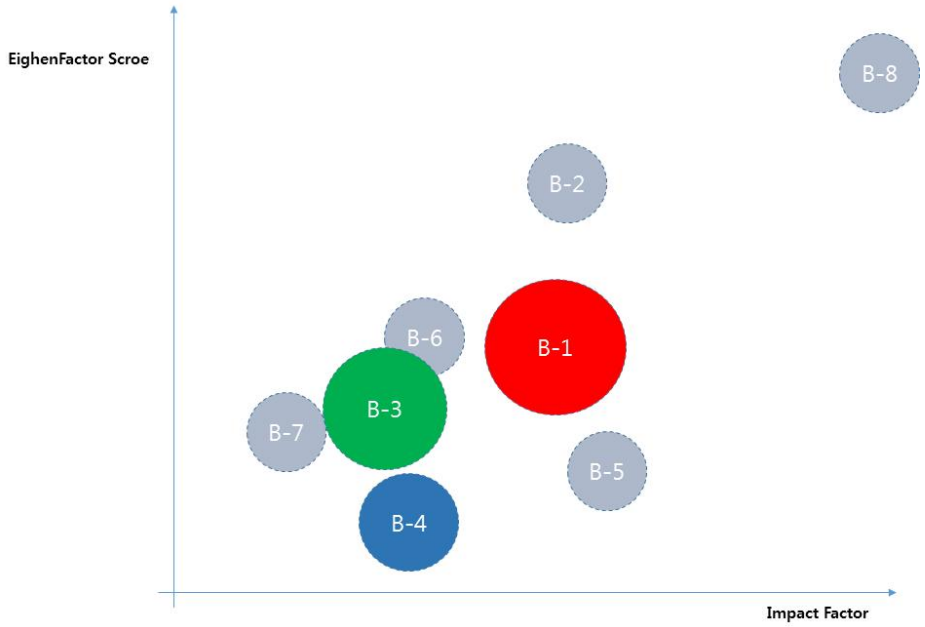
[그림 4-2]는 학술지 (B-1)이 가장 많은 논문을 게재한 학술지라고 가정하였을 때,

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

추천을 하기 위한 카테고리는 해당 저자가 지금까지 써온 전체 학술지가 포함된 카테고리를 선정하여 해당 카테고리 내 학술지와 학술지 (B-1)과 비교하여 이보다 높은 학술지를 추천한다.

[그림 4-3]과 같은 경우 학술지 (B-1)과 (B-8)과의 거리가 매우 멀게 나타난 것을 확인할 수 있다. 하지만 이는 각각 5편, 1편씩 게재한 결과로써 (B-8)을 기준으로 학술지를 추천한다면 게재하는데 문제가 발생할 수 있다. 물론, 현저하게 높은 IF와 ES의 경우가 아닌 학술지라도 한해에 게재한 논문의 양이 낮다면 해당 학술지도 채택율이 낮을 수 있지만 본 논문에서는 IF와 ES를 기준으로 판별하기 때문에 (B-8)를 기준으로 하기에 어려움이 발생한다. 따라서, 학술지 (B-1)을 기준으로 유사 카테고리 내 학술지를 추천한다.

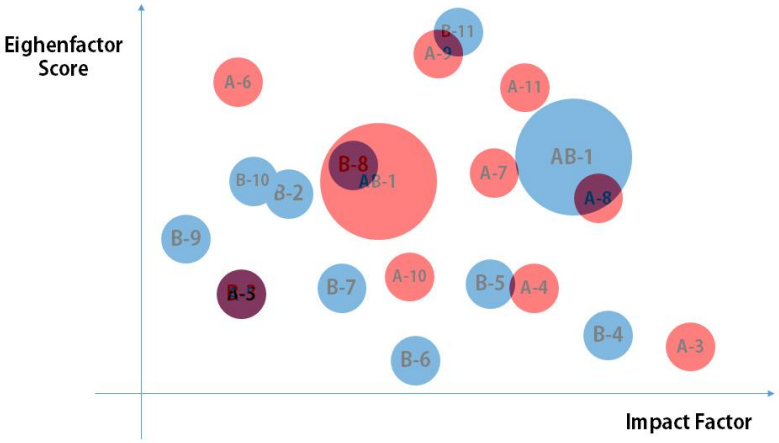


[그림 4-3] Google Scholar 내 연구 실적을 기반으로 한 저널 분포도

마지막으로, 학술지 추천 시 가장 고려되는 부분은 [그림 4-4]와 같다.

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY



[그림 4-4] 특정 학술지 AB-1이 속해있는 두 카테고리 내 학술지 간 거리 측정 예

[그림 4-4]는 본 연구에서 제안하고자 하는 복수의 카테고리에 포함되어 있는 학술지의 경우 각 카테고리 간 거리 측정이 매우 어렵다는 점을 극복하기 위해 방법을 제시하는 것이다. 예를 들어, 특정 학술지 AB-1이 두 카테고리 내 포함되어 있다고 가정할 경우, 두 카테고리 내 포함된 전체 학술지를 EigenFactor Score와 Impact Factor를 기준으로 그래프상에 표현하였을 때, [그림 4-9]와 같다.

먼저, 첫 번째 카테고리(붉은색) 내 존재하는 학술지 간 거리를 보면 AB-1과 가장 가까운 학술지는 A-10이다. 두 번째 카테고리(파란색) 내 존재하는 학술지 간 거리는 B-5와 더 가깝다는 것을 알 수 있다. 하지만, 두 카테고리 내 학술지를 겹쳐놓은 상태에서 가장 가까운 학술지를 선택한다면, 첫 번째 카테고리 경우 B-8, 두 번째 카테고리 경우 A-8이 되는 것을 확인할 수 있다. 따라서, 본 연구에서 제안하는 카테고리 간 유사도 측정 방법을 사용하여 해당 문제를 해결하고자 한다.

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

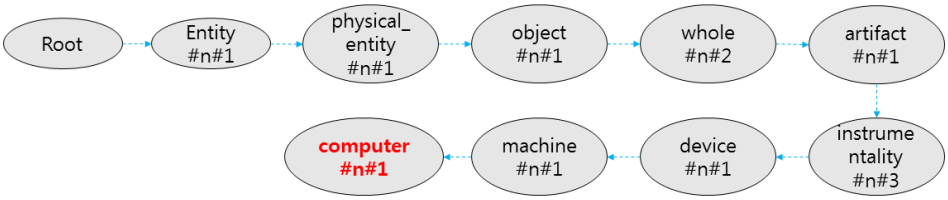
C. EigenValue 기반 카테고리 유사도 측정 방법

1. 카테고리명을 구성하고 있는 단어 간 유사도 측정

본 절에서 카테고리 유사도 측정 방법을 설명한다. 이를 위해 행렬벡터를 사용하여 Eigenvalue 계산 기법을 사용하여 n개의 EigenValue 값을 도출하며, 이를 기하평균 공식을 사용하여 최종적으로 카테고리 유사도를 측정한다.

이를 위해 카테고리 정보를 활용하여 유사도 행렬을 구축해야 한다. 하지만, JCR 정보 내 각각의 카테고리에 추가적인 정보를 추출하는데 한계가 있다는 점을 확인하였다. 이를 보완하기 위해 모든 카테고리명을 선정할 때, 이에 포함된 학술지를 대표할 수 있는 단어들로 구성되어 있다는 점을 확인하였다. 즉, 'Computer Science, Information Systems' 카테고리에는 정보처리 및 시스템 관련 저널들이 대다수 포함되어 있다는 것이다. 예를 들어, 'INFORMATION SCIENCES, INFORMATION & MANAGEMENT, INFORMATION SYSTEMS' 등과 같다.

따라서, 카테고리 간 유사도를 측정하는 방법을 위해 WordNet을 활용한 단어 간 유사도를 측정하였다. WordNet 유사도 측정 방법은 "Wu & Palmer(WUP)", "Leacock & Chodorow(lch)", "Resnik", "Jiang & Conrath", "Lin"이 있다[21, 22, 23, 24, 25]. 이 중 본 논문에서는 단어 간 유사도 측정에 가장 많이 사용되는 "Wu & Palmer(WUP)" 방법을 사용한다. 다른 유사도 측정 방법과 비교 시 LCS(Least Common Subsummer)라는 최소 상위 노드를 기준으로 하여 거리를 측정하는 방법을 사용한다. LCS를 통해 최상위 노드인 'Root'까지의 거리보다 더 작은 값으로 두 단어 사이의 거리를 평준화가 가능하기 때문이다. WordNet의 계층구조를 기반으로 도출된 단어의 계층관계는 [그림 4-5]와 같다.



[그림 4-5] WordNet 기반 'Computer' 단어에 대한 계층구조

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

[그림 4-5]와 같이 최단거리에 두 단어가 공동으로 포함된 가장 가까운 상위노드까지 깊이를 사용한다. 이는 LCS(Least Common Subsummer)에 2를 곱한 후 각각의 두 개의 단어가 공동으로 속한 상위노드까지의 거리를 합한 것으로 나누어 주면 본 연구에서 사용하고자 하는 단어 간 유사도가 구해지게 된다. 이를 통해 두 카테고리명을 기반으로 한 유사도 측정 방법의 결과는 [표 4-6]과 같다. 측정될 유사도 값의 범위는 1~0사이의 값으로 측정되며, 같은 단어일 경우 최대값인 1이 계산되고, POS(Part-of-Speech) 태그정보가 명확하지 않을 때는 에러 값인 -1의 값이 나오고 관련이 없는 경우 0값이 도출된다. 하지만, 본 연구에서의 목적으로 볼 때 오류값으로 표현된 -1값은 모두 관련이 없다는 것으로 간주하여 모두 0값으로 치환하여 계산한다. 앞서 설명한 유사도 측정방법인 "Wu & Palmer" 유사도 측정 방법을 수식화 하면 다음과 같다[26, 27, 28].

$$Sim_{wup}(c_1,c_2) = \frac{2 \times depth(lcs(c_1,c_2))}{depth(c_1)+depth(c_2)}$$ <수식 4-5>

<수식 4-5>를 사용하여 *'Computer Science, Theory Methods'*와 *'Computer Science, Information Systems'* 두 카테고리명을 기준으로 유사도를 측정하면 [표 4-6]과 같다.

[표 4-2] Wu & Palmer Similarity를 활용한 단어 간 유사도 측정 결과

	COMPUTER	SCIENCE	THEORY	METHODS
COMPUTER	1.0000	0.2857	0.2667	0.2667
SCIENCE	0.2857	1.0000	0.7059	0.8000
INFORMATION	0.3077	0.7692	0.7143	0.7143
SYSTEMS	0.7778	0.7500	0.7368	0.9412

2. 각 학술지별 논문 주제(Subject, Abstract, Indexing)를 활용한 유사도 측정

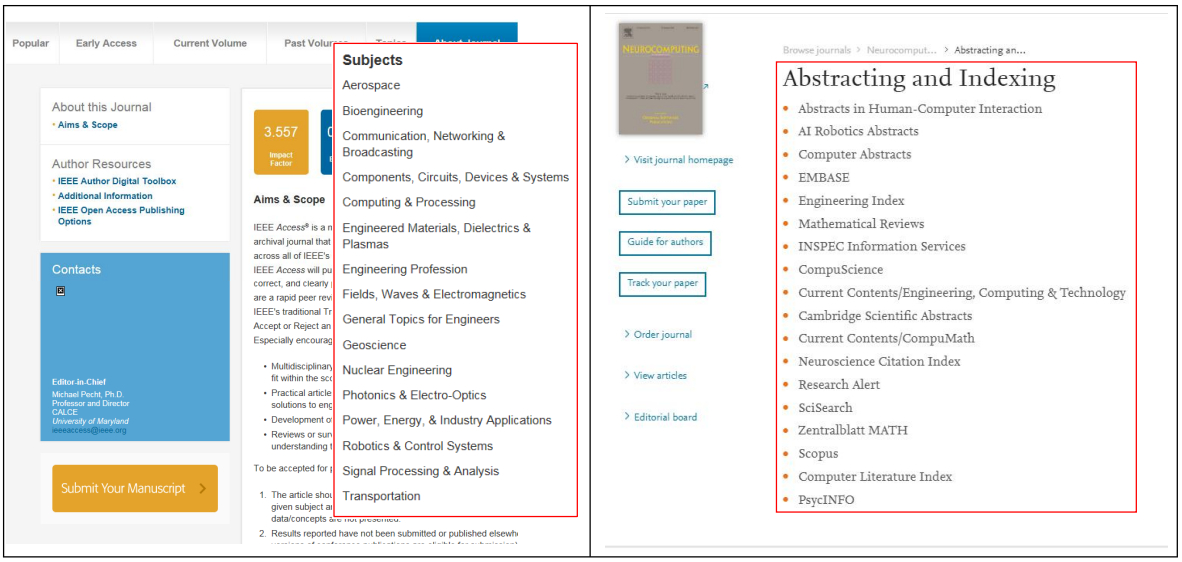
앞서 진행한 카테고리명을 기반으로 한 단어 유사도 측정 방법과 본 논문에서 추가로 사용한 각 학술지별 초록을 사용하고자 한다. 초록이란 논문 또는 보고서 등에서 전체 내용을 요약하여 사용자가 쉽게 이해할 수 있도록 축약해놓은 문단을 말한다. 하

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

지만, 학술지에서의 초록이란 해당 학술지가 취급하고 있는 연구분야 목록을 정리하여 해당 학술지에 제출하고자 하는 사용자에게 알려주는 부분이다. 이를 바탕으로 모든 연구자들은 해당 학술지에 자신의 논문 주제와 연관이 있는지에 대한 여부를 판단하여 논문을 제출한다. 예를 들어, 'IEEE ACCESS' 학술지의 경우 "aerospace, Bioengineering, Communication, Networking & Broadcasting, Components, Components, Circuits, Devices & Systems, Computing & Processing" 등 다수의 학술분야 논문을 다루고 있으며, 'Neurocompuing' 학술지의 경우 "Abstracts in Human-Computer Interaction, AI Robotics Abstracts, Computer Abstracts, EMBASE, Engineering Index, Mathematical Reviews, INSPEC Information Services CompuScience " 등과 같은 분야를 다루고 있다. 이는 [표 4-7]과 같다.

[표 4-3] 'IEEE ACCESS'와 'Neurocomputing' 학술지 내 다루고 있는 학술분야



다루고 위와 같이 특정 학술지에서만 각각의 학술분야를 나타내는 것이 아닌 모든 학술지에서 Subjects, Abstracting and Indexing 과 같은 학술분야를 나타내고 있다. 따라서, 본 연구에서는 해당 부분에 대한 의미분석을 통하여 카테고리 유사도 측정에 사용한다면 의미 있는 데이터로 활용 가능성이 충분하다고 판단하여 학술지별 초록 부분을 수집 분석하였다. 이를 바탕으로 "Wu & Palmer" 유사도 측정 방법을 사용하여 지식베이스를 기반으로 단어 간 유사도측정을 하였다. 이때, 초록 내에 있는 명사만 추출하였으며, 불용어(특수문자)까지 제거하였다. 그 결과 다음과 같은 단어 간 유사

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

도가 추출된다.

[표 4-4] 'IEEE ACCESS'와 'Neurocomputing' 학술지에 포함된 초록을 구성하는 단어 간 유사도 측정결과

ACCESS / Neurocom	Aerospace /NNP	Bioengineering /NNP	Communication /NNP	Systems /NNP	…	Signal /NNP	Processing /NNP	Analysis /NNP	Transportation /NNP
Abstracts/ NNS	0.2667	0.6316	0.6667	0.7368	…	0.6154	0.3077	0.5714	0.4444
Human/NNP	0.3636	0.1600	0.2105	0.4348	…	0.2400	0.3000	0.1905	0.4348
Computer/ NNP	0.4706	0.2353	0.3636	0.7778	…	0.3529	0.5000	0.3077	0.6667
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
Computer/ NNP	0.4706	0.2353	0.3636	0.7778	…	0.3529	0.5000	0.3077	0.6667
Literature/ NNP	0.2857	0.7619	0.7273	0.5714	…	0.6667	0.3333	0.8000	0.6316
Index/NNP	0.3529	0.3333	0.6667	0.7059	…	0.6154	0.4000	0.5882	0.3529

[표 4-8]와 같이 기존 카테고리명을 이루는 단어 간 유사도와 학술지별 초록 내 포함된 단어 간 유사도를 측정한 결과를 바탕으로 카테고리 유사도 측정 방법에 사용된다.

3. 카테고리 유사도 측정 방법

[표 4-6]의 결과를 바탕으로 카테고리 간의 거리값(특성값)을 계산하기 위하여 앞서 설명한 EigenValue를 계산하며, n by n 행렬에 대한 EigenValue값은 총 n개의 결과값을 나타낸다. 먼저 카테고리 유사도 측정 방법에 사용되는 EigenValue 공식에 대한 유도과정을 설명한 후 본 연구에서 사용하는 데이터를 기반으로 카테고리 유사도 측정 방법에 대하여 서술한다.

어떤 행렬 $A = \begin{bmatrix} 1 & 4 \\ 1 & 1 \end{bmatrix}$ 가 있을 때, 이에 대한 EigenValue(λ)를 구하면 <수식 4-6>과 같다.

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

$$\begin{bmatrix}1 & 4\\1 & 1\end{bmatrix}\begin{bmatrix}x_1\\x_2\end{bmatrix}=\lambda\begin{bmatrix}x_1\\x_2\end{bmatrix}$$
$$\left(\begin{bmatrix}1 & 4\\1 & 1\end{bmatrix}-\lambda\right)\begin{bmatrix}x_1\\x_2\end{bmatrix}=0$$
$$\begin{bmatrix}1-\lambda & 4\\1 & 1-\lambda\end{bmatrix}=0$$

<수식 4-6>

위 행렬을 2차방정식으로 풀이하는 과정은 <수식 4-7>과 같다.

$$(1-\lambda)^2-4=0$$
$$\lambda^2-2\lambda-3=0$$
$$(\lambda+1)(\lambda-3)=0$$
$$\therefore \lambda=-1,3$$

<수식 4-7>

위와 같이, 2 by 2 행렬을 예제로 제시하여 도출된 EigenValue값은 2개로써 해당 값은 모두 음수일 수 있으며, 모두 양수일 수 있다. 하지만, 본 논문에서는 음수 또한 중요한 결과치로 생각하여 모든 수에 대해 절대값으로 치환하여 음수는 양수로 치환한다. 또한, 본 논문에서 사용하는 λ값은 한 개의 값을 필요로 하기 때문에 기하평균을 사용하여 이를 계산한다. 하지만, 평균값을 구하는 방법은 크게 산술평균, 기하평균, 조화평균 방법을 사용하는데 본 논문에서 기하평균을 사용한 이유를 계속해서 설명한다. [표 4-5]는 [표4-2]를 기반으로 EigenValue(λ)을 도출한 결과를 나타낸다.

[표 4-5] EigenValue(λ)값 계산 결과

	COMPUTER	SCIENCE	THEORY	METHODS
COMPUTER	2.6190	0	0	0
SCIENCE	0	0.7575	0	0
INFORMATION	0	0	0.1901	0
SYSTEMS	0	0	0	0.0888

최종적으로 도출된 4개의 EigenValue를 사용하여 카테고리 유사도로 사용한다. 일반적으로 위와 같이 4개의 값이 계산 되었을 경우 EigenValue에 대한 최대값을 많이 사용하는데 본 연구에서는 최종적으로 도출된 계산 결과값은 각각의 단어에 대한 특성값이라고 생각하여 기하평균에 의해 최종 값을 구한다. 본 연구에서 사용하는 평균 계산 방법은 기하평균을 사용하는데 이유는 다음 실험 결과와 같다.

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

[표 4-6] 평균값을 구하기 EigenValue 계산 결과 예제

	Word 1	Word 2	Word 3	Word 4
Word 5	9.5414	0	0	0
Word 6	0	0.7575	0	0
Word 7	0	0	0.1901	0
Word 8	0	0	0	0.0888

[표 4-6]은 평균값 계산에 사용되는 산술평균, 조화평균, 기하평균에 대한 결과를 확인해보기 위한 실험데이터로 사용하기 위해 특정값 한 개에 대한 수치를 변경하여 각각의 평균값에 대한 결과를 알아보기 위해 제시한 예제이다.

각각의 평균값을 구하기 위해 [표 4-5], [표 4-6]의 값을 통해 산술, 조화, 기하평균을 구해보면 다음과 같다.

[표 4-7] Lamda A, B에 대한 각각의 평균값

※산술평균 $\frac{2.6190+0.7575+0.1901+0.0888}{4}=0.9139$	※산술평균 $\frac{9.5414+0.7575+0.1901+0.0888}{4}=2.6445$
※조화평균 $\frac{4(2.6190\times0.7575\times0.1901\times0.0888)}{2.6190+0.7575+0.1901+0.0888}=0.1340$	※조화평균 $\frac{4(9.5414\times0.7575\times0.1901\times0.0888)}{9.5414+0.7575+0.1901+0.0888}=0.0461$
※기하평균 $\sqrt[4]{2.6190\times0.7575\times0.1901\times0.0888}=0.2059$	※기하평균 $\sqrt[4]{9.5414\times0.7575\times0.1901\times0.0888}=0.7500$
[표 4-7]에 대한 각각의 평균계산결과	[표 4-8]에 대한 각각의 평균계산결과

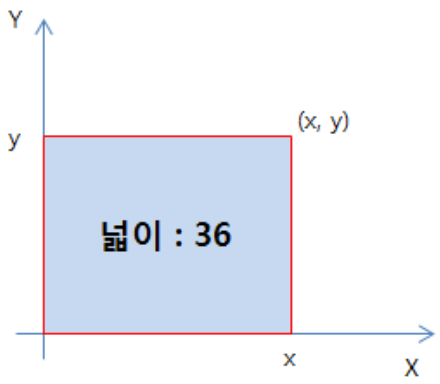
[표 4-7]의 결과를 바탕으로 각각의 평균을 비교해보면, 산술평균의 경우 특정값의 크기에 따라 해당 평균값의 범위 차이가 매우 큰 것을 볼 수 있다. 이와 같이 특이값에 대한 영향을 많이 받게 되면 4개의 값에 대한 정규화를 시키기 힘들다는 것을 확인할 수 있다. 조화평균의 경우 산술평균과 반대로 특정값이 크면 평균값이 오히려 작아지고, 범위가 매우 협소하다는 것을 확인할 수 있다. 이는 추후 최종적으로 계산될 데이터에서 그 범위차이가 작을 경우에 따른 문제점이 발생할 수 있기 때문에 적당한 범위라고 보기힘들다. 조화평균의 경우는 본 논문에서 사용하고자 하는 수치적인 부분에서 결과가 다르기 때문에 이를 사용하지 않는다. 또한, 기하평균의 경우 특정값에 대

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

한 평균값에 영향은 있으나, 산술평균보다는 그 범위가 좁고 조화평균와 반대로 그 값이 커지며 특정값에 따라 범위 차이가 적당하다고 할 수 있다.

하지만, 기하평균을 제외한 산술평균, 조화평균은 실수에서 사용하는 일반적인 평균값을 산출할 수 있으나, 행렬에서는 교환법칙이 성립되지 않기 때문에 이를 정의하는 것은 쉽지 않다[29, 30, 31]. 하지만, Hyun-duk Kim[32]은 행렬 기반 다변수 기하평균에 대해 정의한 행렬평균이론을 바탕으로 본 논문에서 도출한 EigenValue값에 대한 평균을 기하평균을 사용하여 계산한다. 이를 위해 기하평균에 대한 기본적인 의미를 설명한다.



[그림 4-6] 기하평균을 사용한 최적의 좌표값을 구하는 방법에 대한 예

[그림 4-6]과 같이 넓이가 36인 어떤 직사각형이 있을 때, 가장 최적의 x값과 y값은 사각형 넓이 구하는 공식(넓이 = 가로 X 세로)에 의해 다음과 같은 각 좌표점으로 나타낼 수 있다. (1, 36), (2, 18), (3, 12), (4, 9), (6, 6), (9, 4), (12, 3), (18, 2), (36, 1)과 같이 총 9개의 결과를 찾을 수 있는데, 이 중 가장 적절한 좌표값에 대한 평균값은 6으로 나타낼 수 있다. 즉, 특정 좌표점을 이루는 (x, y)값의 범위가 크지 않은 것을 말한다. 하지만, 산술평균의 경우 (1, 36)에 대한 평균을 구하면 18.5, (6, 6)에 대한 평균을 구하면 6이라는 평균을 산출할 수 있다. 이는 각각의 거리 값에 대한 영향을 많이 받는다. 하지만, 기하평균은 두 값을 곱한 결과에 루트($\sqrt{\ }$)를 적용하여 평균값을 도출한다. 즉, (1, 36)에 대한 평균을 구하면 6, (6, 6)에 대한 평균을 구하면 6이라는 평균값을 산출할 수 있다. 따라서, 기하평균은 거리 공간 상에서 특정 수치에 영향을 받지 않고 가장 적합한 평균값을 도출할 수 있다는 내용을 뒷받침

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

한다. 이를 통해 앞서 도출한 EigenValue값을 토대로 기하평균을 사용하여 최종 카테고리 유사도를 측정한 결과는 다음과 같다.

$$Category\,Similarity(C_{CSTM}, C_{CSIS}) = \sqrt[4]{(2.6190 \times 0.7575 \times 0.1901 \times 0.0888} = 2.2059$$

위 근거를 바탕으로 카테고리 유사도 측정 방법을 설명하기 위해 컴퓨터 관련 분야의 카테고리를 활용하여 설명하였으며, 카테고리는 *'COMPUTER SCIENCE, INFORMATION SYSTEMS'*와 *'COMPUTER SCIENCE, THEORY & METHODS'*를 사용하였다.

D. 카테고리 유사도를 이용한 학술지 추천

본 실험에서 사용할 데이터 셋은 *'Computer Science'* 카테고리 내 다양한 저널 중 특정 인물 10명을 무작위로 선택하여 개개인별 연구 결과물을 수집분석한 후 본 연구 방법을 적용하여 연구 분야 분석 및 저널을 추천 결과를 확인하다.

[표 4-8]은 연구자 총 10명에 대한 전체 논문 수 현황과 JCR에 등록된 학술지를 기준으로 재추출한 후 최종 학술지 수를 표기한 것이며, 학술지 수는 중복을 제거하여 나타낸 수이다.

[표 4-8] 저자별 연구 결과물 현황

(단위:편)

저자명(익명)	전체 논문수	JCR 등록된 저널수	저자명(익명)	전체 논문수	JCR 등록된 저널수
Author A (Chang Choi)	9	4	Author F (Jungpil Shin)	18	8
Author B (Ilsun you)	30	22	Author G (Jason J. Jung)	43	32
Author C (Christian Esposito)	16	11	Author H (Florin Pop)	36	16
Author D (Harry Haoxiang wang)	22	12	Author I (Marek R. Ogiela)	25	12
Author E (Aniello Castiglione)	30	22	Author J (Hichem Snoussi)	14	3
전체 합계			10명	243편	77편

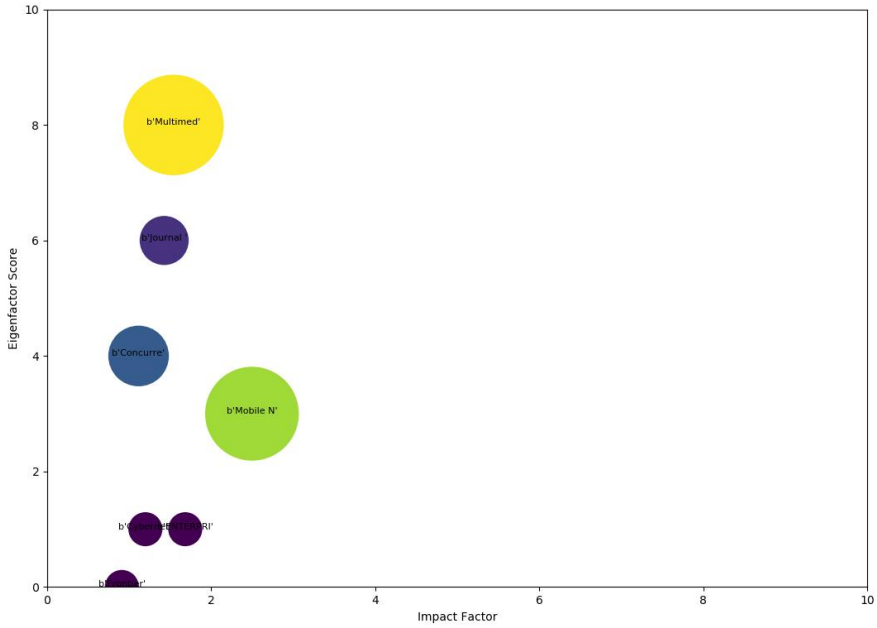
[표 4-8]의 연구결과물을 바탕으로 본 논문에서 제안하는 카테고리 유사도 측정 방

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

법을 적용하여 최종적으로 각 연구자에 대한 최적합 학술지를 추천한다.

연구자별 학술지 간 거리 추출을 위해 JCR에 등록된 학술지 수를 기반으로 Impact Factor와 Eigenfactor Score를 각각 X축과 Y축으로 정의한 후 2차원 평면좌표에 표현하면 [그림 4-7]과 같으며, 이를 통해 각 학술지 간의 거리를 측정한 결과를 바탕으로 현재까지 연구자의 연구 실적을 살펴보면 *"Multimedia Tools and Application"* 이라는 학술지에 가장 많은 논문을 제출한 것을 알 수 있다. 여기서 *"Multimedia Tools and Application"* 학술지는 본 연구에서 정의한 카테고리 포함관계는 총 2곳이 존재한다.



[그림 4-7] Author A의 연구실적 결과를 나타낸 그래프

'Computer Science, Information Systems' 와 *'Computer Science, Software Engineering'* 에 각각 포함되어 있기 때문에 Author A에게 추천해줄 분야는 위의 두 곳을 기준으로 전체 학술지간의 카테고리 간 유사도를 측정한 후 최종적으로 5개의 학술지를 추천하게 된다. 먼저 두 카테고리 간 유사도를 측정한 결과는 [표 4-9]와 같다.

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

[표 4-9] 'Computer Science, Information Systems'와 'Computer Science, Software Engineering' 간 특성값 추출 결과

EigenValue값	기하평균을 사용한 최종 특성값
2.4228 0.1349 0.7361 0.1759	0.1558

[표 4-9]의 카테고리 유사도를 기반으로 *'Multimedia Tools and Application'* 학술지를 기준으로 전체 카테고리 내 포함된 학술지 간 거리를 측정한다. 그 결과는 [표 4-10], [표 4-11]과 같다.

[표 4-10] Author A의 연구 실적 중 *"Multimedia Tools and Applications"*를 기준으로 *"Computer Science, Information Systems"* 카테고리 내 학술지와의 거리값 측정 결과

학술지명	유사도
Journal of Computer Information Systems	0.828452
Cluster Computing The Journal of Networks Software Tools and Applications	0.787596
IT Professional	0.772705
International Journal of Information Security	0.739455
Online Information Review	0.726034
Enterprise Information Systems	0.719857
Multimedia Systems	0.704785
⋮	⋮
IEEE Transactions on Dependable and Secure Computing	0.102665
IEEE Transactions on Services Computing	0.102131
Journal of Information Technology	0.094576
MIS Quarterly	0.048741
IEEE Internet of Things Journal	0.031712
IEEE Network	0
IEEE Transactions on Cloud Computing	0
IEEE Wireless Communications	0
IEEE Communications Surveys and Tutorials	0

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

[표 4-11] Author A의 연구 실적 중 *"Multimedia Tools and Applications"*를 기준으로 *"Computer Science, Software Engineering"* 카테고리 내 학술지와의 거리값 측정 결과

학술지명	유사도
Software Testing Verification & Reliability	0.829422
Software Quality Journal	0.792067
IT Professional	0.772705
IEEE Computer Graphics and Applications	0.754118
International Journal of Information Security	0.739455
Software and Systems Modeling	0.69094
⋮	⋮
Communications of the ACM	0.240711
IEEE Transactions on Visualization and Computer Graphics	0.238366
Advances in Engineering Software	0.220564
IEEE Transactions on Software Engineering	0.202623
IEEE transactions on Multimedia	0.135236

위 [표 4-10], [표 4-11]의 유사도는 앞서 도출한 카테고리 유사도를 학술지 간 유사도에서 값을 빼준 결과값이다. 값을 빼준 이유는 본 연구의 특성 상 최대값을 갖는 학술지 유사도를 상위 5개까지 추천해주기 때문에 수치값이 1을 넘는 경우가 발생한다. 따라서, 유사도 간 가감승을 제외한 제를 하였다. 최종적으로 위의 결과 데이터를 바탕으로 Author A에게 추천해줄 수 있는 카테고리는 두 데이터를 통합하여 나타낸 [표 4-12]를 통하여 확인 할 수 있다.

[표 4-12] Author A에 대한 학술지 추천 목록

순서	학술지명	유사도
1	Software Testing Verification & Reliability	0.829422
2	Journal of Computer Information Systems	0.828452
3	Software Quality Journal	0.792067
4	Cluster Computing The Journal of Networks Software Tools and Applications	0.787596
5	IT Professional	0.772705
6	IEEE Computer Graphics and Applications	0.754118
7	International Journal of Information Security	0.739455
8	Online Information Review	0.726034
⋮	⋮	⋮
251	IEEE Network	0
252	IEEE Transactions on Cloud Computing	0
253	IEEE Wireless Communications	0
254	IEEE Communications Surveys and Tutorials	0

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

[표 4-12]는 각 카테고리별 최상위 논문을 나타내고 있는 [표 4-10]의 *"Journal of Computer Information Systems"* 학술지와 [표 4-11]의 *"Software Testing Verification & Reliability"* 학술지에 대한 두 가지 경우를 상위 5개의 후보리스트에 함께 제시함으로써 한 카테고리만을 고려하여 추천한 결과와는 달리 더욱 폭넓은 학술지 추천이 가능하다. 이는 카테고리 간 유사도 측정방법을 적용하여 기존에 없던 새로운 방식을 통해 한 개의 카테고리가 아닌 다수의 카테고리 내 포함된 학술지까지도 고려하여 추천하였기 때문에 얻을 수 있는 결과이다. 즉, 본 논문에서 제안한 카테고리 유사도라는 연결고리를 통해 다수의 카테고리 간 거리 값을 측정할 수 있는 방법이 사용자의 연구 분야 추천에 대한 새로운 기준점을 마련했다고 볼 수 있다. 이를 바탕으로 추가적인 9명에 대한 논문 추천을 시행한 후 본 논문의 마지막인 부록에 첨부하였다.

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

Ⅴ. 결론 및 향후 연구

A. 결론

본 논문에서는 자신의 연구 결과물을 바탕으로 타 학술지와의 연관정보를 통해 자신의 연구 분야를 쉽게 확인하고, 이를 통해 적절한 학술지를 추천해 줌으로써 연구자들의 연구 능력을 고취시키기 위한 발판을 마련하기 위하여 본 연구를 진행하였다. 이를 위해 카테고리 간의 유사도 측정 방법을 제안함으로써 기존에 수행된 연구에서 시도하지 않은 방법을 적용하였다는 점에 큰 의미가 있다. 카테고리 유사도 측정은 단순 비교가 아닌 지식베이스를 기반한 단어 간 유사도 측정방법을 사용하여 총 2가지의 데이터를 활용하였다.

첫째, 카테고리명에 포함된 단어 간의 유사도를 측정하였으며, 이는 카테고리명 선정 시 해당 카테고리에 포함된 학술지를 모두 포괄할 수 있는 단어로 표현해야 한다는 가정하에 카테고리명을 이루고 있는 단어를 키워드로 선정하였기 때문이다.

둘째로, 각 학술지별 초록 데이터를 활용하였다. 초록은 학술지가 취급하는 학술논문 분야를 포괄적으로 나열해놓은 것으로써 해당 단어들은 학술지를 대표하기 위한 키워드로 판단하였기 때문이다. 이를 토대로 EigenValue값을 산출하여 최종 카테고리 유사도를 판별하였다. 카테고리 유사도는 기존에 없던 카테고리 간의 거리를 측정할 수 있는 하나의 방법으로써 각각 다른 카테고리에 존재하는 학술지까지 고려하여 추천해주기 위해 제안한 것이다.

하지만, 전반적인 논문의 내용은 학술지 내 게재된 논문의 내용을 바탕으로 학술지명과 논문의 내용과 관련성을 평가하기 위한 선행 연구로서 비교평가가 어려웠다는 점이 아쉬운 부분이나, 불명확한 카테고리 간 유사도 측정 방법을 제시함으로써 추가적인 연구가 진행되는데 큰 역할을 할 것으로 기대된다.

이를 토대로 총 10명에 대한 학술지 추천을 진행하였으며, 명확한 비교평가가 어려

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

움에 따라 적절성은 판별하기 힘들지만, 기존에 시도하지 않았던 방법을 적용하고, 단어 간의 의미유사도를 사용하여 학술지를 추천해주었다는 점에서 추후 연구에 꼭 필요한 선행 연구라고 생각된다.

B. 향후 연구

본 연구를 통해 추후 연구의 방향은 본 논문에서 사용한 단어 간 유사도 측정 방법과 관련하여 기타 유사도 측정 방법들과의 비교평가를 통하여 더욱 명확한 유사도 측정 방법을 제시하고자 한다. 또한, 앞서 제안한 평가 방법과 사용자가 제출한 논문들에 대한 키워드 추출 및 주제어를 파악하여 학술지 내 다양한 정보와 매칭하는 방법에 관하여 추가 연구를 진행한다.

또한, 본 연구에서 사용한 각 연구자별 연구 결과 중 최대로 많은 논문을 제출한 학술지뿐만 아니라 특정 빈도수 이상 게재한 다수의 학술지를 바탕으로 추가 데이터를 수집 분석하여 학술지 추천을 위한 방안에 적용을 위한 추가 연구를 하고자 한다.

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

참고문헌

[1] William H. Walters, "Citation-Based Journal Rankings: Key Questions, Metrics, and Data Sources", IEEE ACCESS, Vol. 5, pp. 22036-22053, 2017.

[2] Saeed Iqbal, Muhammad Shaheen, Fazl-e-basit, "A machine learning based method for optimal journal classification", Proceeding of the 8th International Conference for Internet Technology and Secured Transactions, pp. 259-264, 2013.

[3] 조재인, 이지원, "검색포털과 대학도서관 소장목록 연계방안에 관한 연구", 정보관리연구(Journal of information management), Vol. 37, No. 2, pp. 75-92, 2006.

[4] 김진영, "대학생의 포털사이트 학술정보서비스 이용현황 및 개선 방안에 관한 연구", 이화여자대학교 대학원, 석사학위논문, 2012.

[5] Jöran Beel, Bela Gipp, "Google Scholar's Ranking Algorithm: The Impact of Articles' Age (An Empirical Study)", Proceeding of the 2009 Sixth International Conference on Information Technology: New Generations, pp. 160-164, 2009.

[6] Y. Yang, Z. Yanning, "The Effect of Open Access Journals on Citation Impact: A Citation Analysis of Open Access Journals Using Google Scholar", In Proceedings of the In Fourth International Conference on Cooperation and Promotion of Information Resources in Science and Technology, pp. 278-280, 2009.

[7] 박해진, "인터넷 포털사이트 학술정보서비스 이용에 관한 연구", 중앙대학교, 석사학위논문, 2014

[8] Libapps - Online Tutorial - 학술 DB 활용법: Scopus, http://skku.libguides.com/c.php?g=513917&p=3511172, accessed : 29, Sep., 2018

[9] 김용미, "인용학술 데이터베이스 h-지수 비교분석", 석사학위논문, 중앙대학교, 2014

[10] Wikipedia – Impact Factor, https://en.wikipedia.org/wiki/Impact_factor, accessed : 1 Oct., 2018.

[11] 김기홍, "학술지를 평가하는 다양한 지표, 그 특징과 문제점", 과학편집, Vol. 1, No. 1, pp. 34-39.

[12] Wikipedia – Eigenfactor, https://en.wikipedia.org/wiki/Eigenfactor, accessed : 1 Oct., 2018.

[13] 황명권, 김판구, "워드넷의 의미 관계망 증측 방법", 한국정보기술학회논문지, Vol. 7, No. 5, pp. 209-215, 2009.

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

[14] 이은지, “확장된 의미역 결정을 이용한 문서 유사성 판단”, 박사학위논문, 조선대학교 대학원, 2018.

[15] 김형준, “맨하탄 거리망을 위한 멀티 캐스팅 알고리즘”, 석사학위논문, 서울대학교 대학원, 1995.

[16] Y. Oike, M. Ikeda, K. Asada, “A word-parallel digital associative engine with wide search range based on Manhattan distance”, In Proceedings of the IEEE 2004 Custom Integrated Circuits Conference, pp. 295-298, 2004.

[17] Latifa Greche, Maha Jazouli, Najia Es-Sbai, Aicha Majda, Arsalane Zarghili, “Comparison between Euclidean and Manhattan distance measure for facial expressions classification”, In Proceedings of the 2017 International Conference on Wireless Technologies, Embedded and Intelligent Systems, Apr., 2017.

[18] 강성경, 유환, 이영재, “텍스트 마이닝과 소셜 네트워크 분석을 이용한 재난대응 용어 분석”, 한국경영정보학회, Vol. 18, No. 1, pp. 141-155, 2016.

[19] 유광석, 김동욱, 조영송, 김덕수, “평면상에서 원형의 방해물의 존재시 유클리디안 최단경로 계산 알고리듬”, 한국산업경영시스템학회 학술대회, Vol.2003, No.추계, pp. 116-120, 2003

[20] Tistory - 맨하탄 거리(Manhattan Distance)와 유클리디안 거리(Euclidean Distance), http://vsdevelop.tistory.com/2, accessed : 1 Oct., 2018.

[21] Z. Wu, M. Palmer, “Verb semantics and lexical selection,” In Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics, pp. 133-138, 1994.

[22] Claudia Leacock, Martin Chodorow, “Combining local context and wordnet similarity for word sense identification,” WordNet: An Electronic Lexical Database. The MIT Press, pp. 265 - 283, 1998.

[23] Philip Resnik, “Using information content to evaluate semantic similarity,” In Proceedings of the 14th International Joint Conference on Artificial Intelligence, pp. 1-6, 1995.

[24] Jay J. Jiang and David W. Conrath, “Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy,” In Proceedings of International Conference Research on Computational Linguistics, pp. 1-15, 1997.

[25] Dekang Lin, “An Information-Theoretic Definition of Similarity,” Proceedings of the Fifteenth International Conference on Machine Learning, pp. 296-304, 1998.

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

[26] 임초람, 조세형, “워드넷 기반의 단어 중의성 해소 프레임워크”, Journal of Korean Institute of Intelligent Systems, Vol. 23, No. 4, pp. 325-331, 2013.

[27] 강석훈, 박종민, “워드넷을 이용한 문서내에서 단어 사이의 의미적 유사도 측정”, 한국산학기술학회논문지, Vol. 16, No. 11, pp. 7718-7728, 2015.

[28] 최동진, “지능적 문서 분석을 위한 개선된 WSD 방법 연구”, 조선대학교 대학원, 박사학위논문, 2015.

[29] T. Ando, Chi-Kwong Li, Roy Mathias, “Geometric means, Linear Algebra and its Applications”, Vol. 385, pp. 305-334, 2004.

[30] Dario A. Bini, Beatrice Meini, Federico Poloni, “An effective matrix geometric mean satisfying the Ando-Li-Mathias properties”, Mathematics of computation, Vol. 79, pp. 437-452, 2010.

[31] Masahiko Sagae, Kunio Tanabe, “Upper and lower bounds for the arithmetic-geometric-harmonic means of positive definite matrices”, Linear and multi-linear Algebras, Vol. 37, No. 4, pp. 279-282, 1994.

[32] Hyun Duk, Kim, “Geometric mean of Toeplitz Positive Definite Matrices : 토플리치 양정치 행렬들의 기하평균”, 경북대학교 대학원 수학과, 석사학위논문, 2012.

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

부 록

부록은 본 논문에서 제안하는 카테고리 유사도 측정 방법을 적용하여 사용자에게 학술지 추천 결과를 확인하기 위하여 추가로 진행한 것이다. 부록 내용을 확인하기 앞서 확인해야할 사항은 다음과 같다.

Author B, E와 Author D, F, G의 두 그룹에 대한 EigenValue값이 같은 이유는 각 사용자의 연구 결과를 수집하여 분석했을 때, 가장 많은 빈도수를 차지한 학술지가 동일한 카테고리 내 포함되어 있는 경우 최종 유사도 값이 같다는 점 때문에 EigenValue값이 동일한 것이다. 해당 부분에 대하여 추후 연구를 통해 특정 가중치 부여 기법을 적용할 예정이다.

1. Author B에 대한 논문 추천

[부록 1-1] Author B의 전체 연구 실적 정리

학술지명	논문 게재 수
IEEE Access	7
Sensors	2
Journal of Network and Computer Applications	2
Mobile Information Systems	2
Future Generation Computer Systems	1
Computer Networks	1
EURASIP Journal on Wireless Communications and Networking	1
IEEE Network	1
Computational Science and Engineering (CSE) and Embedded and Ubiquitous	1
IEEE Systems Journal	1
PloS one	1
COMPUTER SYSTEMS SCIENCE AND ENGINEERING	1
Information Sciences	1

[부록 1-2] ‘Computer Science, Information Systems’와 ‘Computer Science, Theory & Methods’ 간 특성값 추출 결과

EigenValue값	기하평균을 사용한 최종 특성값
0.0112 0.3357 0.8416 1.4061	0.0034

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

[부록 1-3] Author B에 대한 학술지 추천 목록

순서	학술지명	유사도
1	INTERNATIONAL JOURNAL OF GENERAL SYSTEMS	0.741114
2	IEEE Cloud Computing	0.740844
3	Journal of Grid Computing	0.697845
4	FUZZY SETS AND SYSTEMS	0.684451
5	COMPUTER METHODS AND PROGRAMS IN BIOMEDICINE	0.661641
⋮	⋮	⋮
251	IEEE TRANSACTIONS ON MOBILE COMPUTING	0.411312
252	IEEE TRANSACTIONS ON MULTIMEDIA	0.405484
253	Journal of Cheminformatics	0.312041
254	INFORMATION & MANAGEMENT	0.151456

2. Author C에 대한 논문 추천

[부록 2-1] Author C의 전체 연구 실적 정리

학술지명	논문 게재 수
IEEE Communications Magazine	3
Future Generation Computer Systems	3
IEEE Cloud Computing	2
Concurrency and Computation: Practice and Experience	1
Pervasive and Mobile Computing	1
Information Sciences	1

[부록 2-2] 'Computer Science, HardWare & Architecture'와 'Computer Science, Artificial Intelligence' 간 특성값 추출 결과

EigenValue값	기하평균을 사용한 최종 특성값
0.0479 0 0.7574 1.9904	0.0456

[부록 2-3] Author C에 대한 학술지 추천 목록

순서	학술지명	유사도
1	MICROPROCESSORS AND MICROSYSTEMS	0.784631
2	ACM Transactions on Reconfigurable Technology and Systems	0.764151
3	CANADIAN JOURNAL OF ELECTRICAL AND COMPUTER ENGINEERING-REVUE CANADIENNE DE GENIE ELECTRIQUE ET INFORMATIQUE	0.760451
4	IBM JOURNAL OF RESEARCH AND DEVELOPMENT	0.758161

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

5	ACM Transactions on Storage	0.751546
⋮	⋮	⋮
194	ARTIFICIAL INTELLIGENCE REVIEW	0.369154
195	EXPERT SYSTEMS WITH APPLICATIONS	0.271516
196	INTEGRATED COMPUTER-AIDED ENGINEERING	0.148412
197	JOURNAL OF INTELLIGENT MANUFACTURING	0

3. Author D에 대한 논문 추천

[부록 3-1] Author D의 전체 연구 실적 정리

학술지명	논문 게재 수
Cluster computing	8
Sustainable Computing: Informatics and Systems	2
Pattern Recognition Letters	1
Computers & Electrical Engineering	1

[부록 3-2] 'Computer Science, Information Systems'와 Computer Science, Software Engineering 간 특성값 추출 결과

EigenValue값	기하평균을 사용한 최종 특성값
2.4228 0.1349 0.7361 0.1759	0.1558

[부록 3-3] Author D에 대한 학술지 추천 목록

순서	학술지명	유사도
1	Journal of Computer Information Systems	0.945165
2	International Journal of Information Security	0.915461
3	Software Quality Journal	0.904421
4	Cluster Computing The Journal of Networks Software Tools and Applications	0.846152
5	Software Testing Verification & Reliability	0.813144
⋮	⋮	⋮
251	IEEE Computer Graphics and Applications	0.314574
252	Online Information Review	0.241533
253	IEEE Network	0.148158

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

4. Author E에 대한 논문 추천

[부록 4-1] Author E의 전체 연구 실적 정리

학술지명	논문 게재 수
Future Generation Computer Systems	4
IEEE Cloud Computing	4
IEEE Communications Magazine	3
Information Sciences	3
ACM Transactions on Embedded Computing Systems	1
Computers & Security	1
Concurrency and Computation: Practice and Experience	1
IEEE Systems Journal	1
IEEE Transactions on Emerging Topics in Computing	1
IEEE Transactions on Sustainable Computing	1
International Journal of Foundations of Computer Science	1
Pattern Recognition Letters	1

[부록 4-2] 'Computer Science, Information Systems'와 'Computer Science, Theory & Methods' 간 특성값 추출 결과

EigenValue값	기하평균을 사용한 최종 특성값
0.0112 0.3357 0.8416 1.4061	0.0034

[부록 4-3] Author E에 대한 학술지 추천 목록

순서	학술지명	유사도
1	IEEE Transactions on Cloud Computing	0.888413
2	Future Generation Computer Systems-The International Journal of eScience	0.864811
3	Information Fusion	0.854113
4	IEEE Transactions on Information Forensics and Security	0.841354
5	ACM COMPUTING SURVEYS	0.840155
⋮		⋮
148	IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS	0.516413
149	MULTIMEDIA SYSTEMS	0.451351
150	International Journal of Information Security	0.413252
151	Swarm and Evolutionary Computation	0.401361
152	HUMAN-COMPUTER INTERACTION	0.384612

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

5. Author F에 대한 논문 추천

[부록 5-1] Author F의 전체 연구 실적 정리

학술지명	논문 게재 수
Journal of Internet Technology	3
IEEE Access	2
Cluster Computing	1
International Journal of Foundations of Computer Science	1
Journal of Advanced Research in Dynamical and Control Systems	1

[부록 5-2] 'Computer Science, Information Systems'와 Computer Science, Software Engineering 간 특성값 추출 결과

EigenValue값	기하평균을 사용한 최종 특성값
2.4228 0.1349 0.7361 0.1759	0.1558

[부록 5-3] Author F에 대한 학술지 추천 목록

순서	학술지명	유사도
1	JOURNAL OF INFORMATION SCIENCE	0.874612
2	ACM Transactions on Internet Technology	0.861542
3	Cluster Computing-The Journal of Networks Software Tools and Applications	0.854613
4	Journal of Internet Technology	0.814641
5	JOURNAL OF INTELLIGENT INFORMATION SYSTEMS	0.791513
⋮		⋮
224	IEEE Transactions on Affective Computing	0.461346
225	HUMAN-COMPUTER INTERACTION	0.301514
226	USER MODELING AND USER-ADAPTED INTERACTION	0.254942
227	CYBERNETICS AND SYSTEMS	0.154610
228	Universal Access in the Information Society	0.143332

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

6. Author G에 대한 논문 추천

[부록 6-1] Author G의 전체 연구 실적 정리

학술지명	논문 게재 수
Multimedia Tools and Applications	8
Mobile Networks and Applications	7
Future Generation Computer Systems	5
IEEE Access	3
Journal of Intelligent & Fuzzy Systems	2
Concurrency and Computation: Practice and Experience	3
Cybernetics and Systems	1
ENTERPRISE INFORMATION SYSTEMS	1
Expert Systems with Applications	1
Frontiers of Information Technology & Electronic Engineering	1

[부록 6-2] 'Computer Science, Information Systems'와 Computer Science, Software Engineering 간 특성값 추출 결과

EigenValue값	기하평균을 사용한 최종 특성값
2.4228 0.1349 0.7361 0.1759	0.1558

[부록 6-3] Author G에 대한 학술지 추천 목록

순서	학술지명	유사도
1	Software Testing Verification & Reliability	0.844413
2	Journal of Computer Information Systems	0.843475
3	Software Quality Journal	0.781414
4	Cluster Computing The Journal of Networks Software Tools and Applications	0.780145
5	IT Professional	0.751642
6	IEEE Computer Graphics and Applications	0.741645
7	International Journal of Information Security	0.734561
8	Online Information Review	0.715461
⋮	⋮	⋮
251	IEEE Network	0
252	IEEE Transactions on Cloud Computing	0
253	IEEE Wireless Communications	0
254	IEEE Communications Surveys and Tutorials	0

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

7. Author H에 대한 논문 추천

[부록 7-1] Author H의 전체 연구 실적 정리

학술지명	논문 게재 수
Future Generation Computer Systems	4
IEEE Communications Magazine	2
Soft Computing	2
Concurrency and Computation: Practice and Experience	1
IEEE Cloud Computing	1
Information Sciences	1
International Journal of Applied Mathematics and computer Science	1
Journal of Computational Science	1
Journal of Systems Architecture	1
Microprocessors and Microsystems	1
STUDIES IN INFORMATICS AND CONTROL	1

[부록 7-2] 'Computer Science, Information Systems'와 Computer Science, Software Engineering 간 특성값 추출 결과

EigenValue값	기하평균을 사용한 최종 특성값
2.4228 0.1349 0.7361 0.1759	0.1558

[부록 7-3] Author H에 대한 학술지 추천 목록

순서	학술지명	유사도
1	IEEE Transactions on Cloud Computing	0.913542
2	Future Generation Computer Systems-The International Journal of eScience	0.897741
3	Information Fusion	0.891543
4	IEEE Transactions on Information Forensics and Security	0.814654
5	ACM COMPUTING SURVEYS	0.791546
⋮	⋮	⋮
211	IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS	0.246121
212	MULTIMEDIA SYSTEMS	0.201451
213	International Journal of Information Security	0.195462
214	Swarm and Evolutionary Computation	0.146513
215	HUMAN-COMPUTER INTERACTION	0.105714

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

8. Author I에 대한 논문 추천

[부록 8-1] Author I의 전체 연구 실적 정리

학술지명	논문 게재 수
Computational Intelligence (SSCI)	2
Future Generation Computer Systems	2
Optics and Photonics for Information Processing	2
IEEE Systems Journal	1
International Journal of Information Management	1
Journal of Computer and System Sciences	1
Multimedia Tools and Applications	1
przeglad Elektrotechniczny	1
Sensors	1

[부록 8-2] 'Computer Science, Interdisciplinary Applications'와 'Computer Science, Artificial Intelligence' 간 특성값 추출 결과

EigenValue값	기하평균을 사용한 최종 특성값
0 0.1832 0.7234 1.8300	0.1541

[부록 8-3] Author I에 대한 학술지 추천 목록

순서	학술지명	유사도
1	JOURNAL OF INTELLIGENT MANUFACTURING	0.874561
2	HUMAN-COMPUTER INTERACTION	0.845561
3	International Journal of Information Security	0.834546
4	INTERNATIONAL JOURNAL OF INTELLIGENT SYSTEMS	0.791543
5	ACM COMPUTING SURVEYS	0.746621
⋮		⋮
243	IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS	0.045941
244	MULTIMEDIA SYSTEMS	0
245	Cognitive Computation	0
246	Swarm and Evolutionary Computation	0
247	DECISION SUPPORT SYSTEMS	0

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

9. Author J에 대한 논문 추천

[부록 9-1] Author J의 전체 연구 실적 정리

학술지명	논문 게재 수
Engineering Applications of Artificial Intelligence	2
IEEE Access	1

[부록 9-2] 'Computer Science, Artificial Intelligence'와 Computer Science, Software Engineering' 간 특성값 추출 결과

EigenValue값	기하평균을 사용한 최종 특성값
0.0468 0 0.7631 1.9837	0.0423

[부록 9-3] Author J에 대한 학술지 추천 목록

순서	학술지명	유사도
1	International Journal of Computational Intelligence Systems	0.94521
2	INTERNATIONAL JOURNAL OF INFORMATION TECHNOLOGY & DECISION MAKING	0.91546
3	JOURNAL OF INTELLIGENT & FUZZY SYSTEMS	0.87456
4	IEEE INTELLIGENT SYSTEMS	0.81134
5	KNOWLEDGE AND INFORMATION SYSTEMS	0.75662
⋮	⋮	⋮
204	IEEE TRANSACTIONS ON MEDICAL IMAGING	0.03485
206	COMPUTER-AIDED CIVIL AND INFRASTRUCTURE ENGINEERING	0.02115
207	IEEE Transactions on Industrial Informatics	0
208	MEDICAL IMAGE ANALYSIS	0

Collection @ chosun