정보 검색에서의 잠재 의미 분석 방법을 이용한 응집 계층 군집화 기법 연구

Abdel-Ilah Zakaria Khiati*, 강대현*, 박한샘*, 권경락*, 정인정*
*고려대학교 컴퓨터정보학과
e-mail : {le_zakkaz, internetkbs, park1123000, helpnara, chung}@korea.ac.kr

Agglomerative Hierarchical Clustering Using Latent Semantic Analysis in Information Retrieval

Abdel-Ilah Zakaria Khiati*, Daehyun Kang*, Hansaem Park*, Kyunglag Kwon*, In-Jeong Chung*
*고려대학교 컴퓨터정보학과

요 약

본 논문에서는 정보 검색 분야에서 잘 알려진 잠재 의미 분석 방법과 계층적 군집화 방법의 단점을 상호 보완하여 보다 효율적인 정보 검색을 위한 혼합형 군집화 방법을 제안한다. 먼저, 잠재 의미 분석 방법은 벡터 연산을 통하여 자동적으로 문서 내에 있는 잠재적인 의미를 찾는 정보 검색 분야에서 많이 사용되는 고전적인 방법이다. 그러나 이 방법은 언어의 유의성이나 다의성으로 인하여 발생되는 백-오브-워드(bag-of-word) 문제를 가지고 있다. 두 번째 방법인 문서 군집화를 위하여 범용적으로 사용되고 있는 계층적 군집화 방법이다. 이 방법은 이를 통하여 분석된 군집의 질적 측면에서 볼 때, 여전히 단층적 군집들이 많이 형성되어 세부적인 분석을 통한 추가적인 군집화가 필요함을 알 수 있다. 따라서, 본 논문에서는 앞서 언급한 문제점을 해결하기 위하여 혼합적인 방법으로 잠재 의미 분석 방법을 이용한 응집 계층 군집화 방법을 제안한다. 제안한 방법을 이용하여 잘 알려진 두 개의 데이터에 적용하고 기존의 방법과 그 결과를 비교함으로써 군집의 질적 측면에서의 우수함을 보인다.

1. 서론

최근 웹을 통하여 다양하고 방대한 양의 정보가 생성되고 공유됨에 따라 사용자가 원하는 정보를 검색하는데 있어서 많은 어려움이 있다. 특히, 동의어나 다의어들로 인하여, 사용자가 원하는 최종 결과를 검색하기 위하여 사용자는 검색 엔진으로부터 추출된 검색 결과를 다시 검토하거나 분류해야 한다. 이 때, 추출된 검색 결과들은 많은 양의 스니핏(snippet)들로 구성되어 있다. 그림 1 은 구글 검색 질의 창에서 재규어라는 질의어(query)를 입력하였을 때 검색된 검색 결과 목록에서 추출된 스니핏의 예제이다.


재규어 - 위키백과, 우리 모두의 백과사전
ko.wikipedia.org/wiki/재규어 ▾ Translate this page Korean Wikipedia ▾
재규어(jaguar, 학명: Panthera onca)는 신대륙에서 사는 고양이과 척추동물로, 구대륙의 호랑이, 사자, 표범과 함께 표범속의 네 '큰 고양이' 중 하나이다. 호랑이, 사자 ...
생태 - 어원 - 분류 - 분포 및 서식지

(그림 1) 구글에서 검색된 스니핏의 예제

스니핏은 해당 사이트의 제목, URL 주소와 사이트에 대한 간략한 설명으로 구성되어 사용자가 원하는 정보를 찾을 수 있도록 도와준다. 그러나, 이러한 스니핏들을 분류의 구분이 없이 모두 섞여서 추출되기 때문에 사용자는 이를 하나하나 검토해야 하는 불편함이 발생한다. 예를 들어, “재규어”라는 단어를 검색 질의 창에 입력하면 자동차 브랜드인 재규어, 큰 고양이과 중 하나인 재규어, 만화나 음악 주제인 재규어 등 여러 가지 의미를 가진 동의어들이 섞여서 검색 결과에 나타난다. 이러한 검색 결과의 질을 높이기 위하여 서픽스 트리 군집화(Suffix Tree Clustering, STC [1]), 링고 (Lingo [2]) 등과 같은 다양한 방법이 연구되고 있다. 그리고 웹 기반 시스템으로는 Carrot2 [3], Vivisimo (Clusty, Yippy) [4] 등이 있다.

그러나, 기존의 연구 방법에서도 언어의 유의성이나 다의성으로 인하여 단어들 사이의 의미를 고려하지 못하는 백-오프-워드 문제를 가지고 있다[5]. 뿐만 아니라, K-means 알고리즘과 같은 비계층적 군집화 방법을 이용하여 이러한 문제를 해결하고자 다양한 연구가 진행되었으나, 최하위 노드에서의 단층 군집(flat cluster)들이 형성됨으로 인하여 올바르게 분류하지 못하는 한계를 가지고 있다[1-4].

따라서 본 논문에서는 위의 언급된 문제점들을 해결하고 상호 보완하기 위하여 혼합적인 방법으로 잠재 의미 분석 방법을 이용한 응집 계층 군집화 방법을 제안한다. 계층적 군집화 방법을 이용함으로써 비계층적 군집화 방법이 가지고 있는 군집의 개수 결정, 비결정적이며 비구조적인 군집화, 군집화된 정보의 가독성 등의 문제들을 해결할 수 있다. 그리고 잠재

의미 분석 기법을 이용함으로써 다른 주제간의 군집을 최소화시킬 수 있으며, 추가적인 군집이 필요한 군집에 대해서도 다시 군집화를 통하여 더욱 군집의 질을 높일 수 있다.

실험을 통하여 제안한 방법의 효율성과 군집들의 질적 향상을 보이며, 정확도, 재현율, F-측정과 같은 평가 척도를 이용하여 성능을 검증한다. 제안한 방법을 기존의 잘 알려진 AMBIENT 와 MORESQUE 데이터 집합에 적용하고 기존의 방법과 비교하고 군집의 질적 향상을 보인다.

2. 관련 연구

2.1. 잠재 의미 분석 (Latent Semantic Analysis, LSA)

정보 검색 분야에서의 잠재 의미 분석 방법(LSA)은 문서의 색인화 및 검색, 문서 차원 축소 등의 다양한 분야에 활용된다[6]. 특히, 이 방법은 수학적 방법인 특이값 분해 (Singular Value Decomposition, SVD)를 이용하여 단어나 문서들 사이의 의미론적 관계를 분석하고 발견한다. 즉, 같은 주제에 있는 문서들은 공통된 단어를 사용한다는 점에 기인한 것이다. 기존 단어 기반 정보를 표현하는 벡터 공간 모델(Vector Space Model)과 유사하지만, 벡터 공간 모델이 가지고 있는 유의성이나 다의성으로 인하여 나타나는 백-오브-워드(bag-of-word) 한계를 해결하기 위해 고안되었다[7, 8]. LSA 는 키워드 기반이 아닌 주제를 기반으로 문서를 탐색하기 때문에 연관되지 않은 키워드를 가진 문서는 결과에 포함되지 않는다. 이러한 SVD 는 다음과 같이 표현된다.

$$A = S\Sigma U^T \quad \text{(식 1)}$$

식 1 에서 $\Sigma = \text{diagonal}(a_1, a_2, \cdots, a_n)$ and $a_1 > a_2 > \cdots > a_n$를 나타내며, 각 원소는 행렬 A의 특이값들로 구성된다. 행렬 S와 U는 각각 행렬 A의 좌, 우측 특이값을 나타낸다. 첨자 T는 전치행렬(transpose matrix)을 의미한다. 문서들에서 단어들과 문서들은 각각 의미 공간(Semantic Space)인 $S\Sigma$와 ΣU^T으로 표현된다.

본 논문에서는 웹 상의 빅데이터 처리 및 행렬의 차원 축소를 위하여 다양한 특이값 분해[9, 10] 방법 중 Truncated 특이값 분해 방법을 이용하여 잠재 의미 분석을 수행한다. 이는 다음과 같이 표현된다.

$$\tilde{A} = S_k\Sigma_k U_k^T \quad \text{(식 2)}$$

식 2 에서 행렬Σ에서 k번째로 큰 특이값들로 구성된 행렬Σ_k를 생성한 후, k열로 구성된 행렬을 S_k로 나태내고, k행으로 구성된 행렬을 U_k^T로 표현한다.

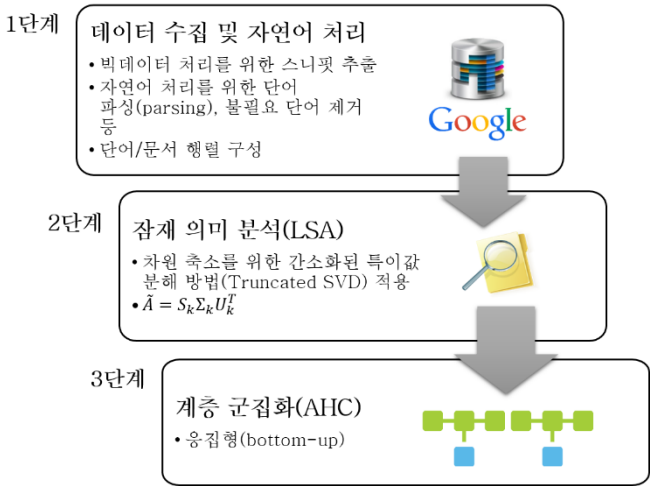
2.2. 계층 군집화 (Hierarchical Clustering)

계층적 군집화 방법은 군집화 방법에 따라 크게 상향식(응집, bottom-up)과 하향식(top-down)으로 나뉜다. 상향식 방법은 각 문서를 하나의 단일 군집으로 간주하고 시작하여 하나의 군집으로 형성될 때까지 반복하여 군집화를 하는 방식이다. 상향식 방법과 반대의 방법인 하향식 방법은 전체 문서를 하나의 군집으로 간주한 후, 이를 조건에 따라 두 개 이상의 새로운 군집으로 나누는 방식이다. 이는 각 각의 문서가 하나의 군집에 포함될 때까지 반복한다. 이 방법에서 사용되는 새로운 군집을 생성하는 조건은 다루고자 하는 데이터의 형태에 따라 다양하다.

두 방법은 모두 계층적 구조의 군집 결과를 나타낸다. 일반적으로 계층적 군집 방법은 $O(n^3)$의 복잡도를 가지기 때문에[11], 방대한 양의 빅데이터 처리에는 부적합하지만, 본 논문에서는 문서 전체를 대상으로 하지 않고 검색 결과에서 나타난 스니핏만을 사용하기 때문에 복잡도로 인한 문제점을 해결할 수 있다.

3. LSA 를 이용한 응집 계층 군집화 기법

그림 2 와 같이 제안한 방법은 1) 데이터 수집 및 자연어 처리, 2) 잠재 의미 분석, 3) 응집 계층 군집화, 크게 3 단계로 구성되어 있다. 각 단계에 대한 설명은 다음 절에서 자세하게 설명한다.



(그림 2) 제안한 방법의 전체 구성도

3.1. 데이터 수집 및 자연어 처리

본 논문에서 사용되는 데이터는 구글이나 야후 등과 같은 검색엔진에서 질의를 하였을 때 나오는 검색 결과인 스니핏들이다. 그림 1 과 같이 각 스니핏은 사이트의 URL, 제목, 간략한 설명으로 구성되어 있다. 수집된 스니핏을 질의에 따라 분류한 후, 각 스니핏에 대해서 파싱(parsing), 불필요한 단어 제거(예: a, an, the 등), 단어 원형 복원(예: computing → compute) 등의 자연어 처리를 위한 정제 작업을 수행한다. 본 논문에서 사용되는 데이터는 검색 결과로부터 추출된 스니핏과 이를 자연어 처리를 통하여 분석한 단어들이기 때문에, 다음 단계의 잠재 의미 분석 기법과 응집 계층 군집화 알고리즘을 혼합하여 적용하는 경우라도 빠른 시간 내에 결과를 추출할 수 있다.

3.2. 잠재 의미 분석

2 장에서 언급한 언어의 다의성, 유의성 문제를 해결하기 위하여 이전 단계로부터 추출된 단어들에 대

하여 잠재 의미 분석 기법을 적용한다. 이 기법은 단어 빈도수 행렬의 가중치와 행렬의 차원(k), 유사도 측정 방법에 따라 분석 결과가 다르게 나타난다.

먼저 단어 빈도수 행렬의 가중치(weight)를 부여하기 위한 방법으로는 로그 엔트로피(log entropy), TF-IDF (Term Frequency-Inverted Document Frequency) 가중치 방법 등이 있는데, 본 논문에서는 가장 널리 사용되는 TF-IDF 가중치 방법을 사용한다.

$$w_{ij} = \log(tf_{ij} + 1) \times \log\frac{N}{idf_j} \quad \text{(식 3)}$$

식 3 에서 N은 전체 문서의 수, tf_{ij}는 문서 i내에 있는 단어 j의 빈도수, idf_j는 단어 j를 포함하고 있는 문서의 수를 나타낸다.

가중치 함수가 결정되면 행렬의 차원을 줄이기 위해 특이값 분해(SVD) 방법을 사용하는데 이 때 필요한 차원의 크기(k)를 결정하기 위하여 본 논문에서는 많은 연구에서 활용되는 다음의 프로베니우스 놈(Frobenius Norm)을 사용한다 [12].

$$k = \sqrt{\sum_{i=1}^{\min\{m,n\}} |\delta_i^2|} \quad \text{(식 4)}$$

식 4 에서 m과 n은 각각 단어와 문서의 수를 나타내며, δ_i는 행렬 Σ에서 특이값들을 의미한다. 끝으로 $S\Sigma$와 ΣU^T의 유사도(similarity)를 측정한다. 유사도 측정을 위한 방법으로는 코사인(Cosine) 거리 방법, 유클리디안(Euclidean) 거리 방법, 피어슨(Pearson) 거리 방법 등이 있다. 본 논문에서는 두 벡터 $\boldsymbol{a}$와 $\boldsymbol{b}$의 유사도$_{\boldsymbol{a},\boldsymbol{b}}$를 측정하기 위해 다음 수식과 같이 코사인 거리 측정 방법을 이용한다.

$$\text{유사도}_{\boldsymbol{a},\boldsymbol{b}} = \frac{\sum_{i=1}^{n}(\boldsymbol{a}_i \times \boldsymbol{b}_i)}{\|\boldsymbol{a}\| \cdot \|\boldsymbol{b}\|} \quad \text{(식 5)}$$

식 5 에서 i는 각 벡터에서의 위치를 나타내며, $\|\boldsymbol{a}\|$와 $\|\boldsymbol{b}\|$은 두 벡터 $\boldsymbol{a}$와 $\boldsymbol{b}$의 각각 유클리디안 놈(norm)을 의미한다.

3.3. 응집 계층 군집화

1 장에서 언급한 군집화에서 발생하는 단층 군집 형성과 군집의 개수 결정 문제를 해결하기 위하여 스니핏들로부터 잠재 의미 분석 기법을 통해 생성된 행렬에 응집 계층 군집화 방법을 적용한다.

응집 계층 군집화 방법은 다음과 같이 반복적인 단계에 의해 수행된다. 먼저, n개의 스니핏$(s_i, 1 \le i \le n)$에 대하여 잠재 의미 분석 기법을 이용하여 분석된 결과를 n 개의 군집 $\left(C_0 = \{c_1 = \{s_1\}, c_2 = \{s_2\}, \cdots, c_n = \{s_n\}\}\right)$으로 구성한다. 모든 $i \neq j$ 조건을 만족하는 군집쌍(c_i, c_j)에 대해서 가장 높은 유사성을 가진 두 군집을 하나의 군집으로 합친다. 합쳐진 두 군집을 제거하고 이를 새로운 군집으로 추가한다. 이 과정을 모든 군집이 하나의 군집으로 합쳐질 때까지 수행한다. 이렇게 수행한 결과는 트리(tree)로 표현된다. 이 트리의 최상위 노드(root node)는 하나의 군집으로 표현되고, 최하위 노드(leaf node)들은 하나의 스니핏으로 구성된 n개의 군집들로 표현된다.

4. 실험 및 평가

4.1. 데이터 및 평가 기준

제안한 방법의 성능 평가를 위하여 본 논문에서는 표 1 과 같이 두 가지의 알려진 데이터를 이용한다.

<표 1> 실험에 이용된 데이터

구 분	AMBIENT [13]	MORESQUE [14]
검색 결과 출처	위키피디아 야후 검색엔진	야후 검색엔진
주제 수	44	114 + 하위 주제
주제당 스니핏 수	100	상위 100

제안한 방법의 성능 검증을 위해 위의 두 데이터를 이용하여 실험을 수행한다. 제안된 방법의 성능 검증을 위하여 다음의 식 6, 7, 8, 9 와 같이 정보 검색 분야에서의 대표적인 평가 척도인 정확도, 재현율, F-measure (F), 랜드 인덱스(Rand Index, RI) [15]를 이용한다.

$$\text{정확도} = tp/(tp + fp) \quad \text{(식 6)}$$

$$\text{재현율} = tp/(tp + fn) \quad \text{(식 7)}$$

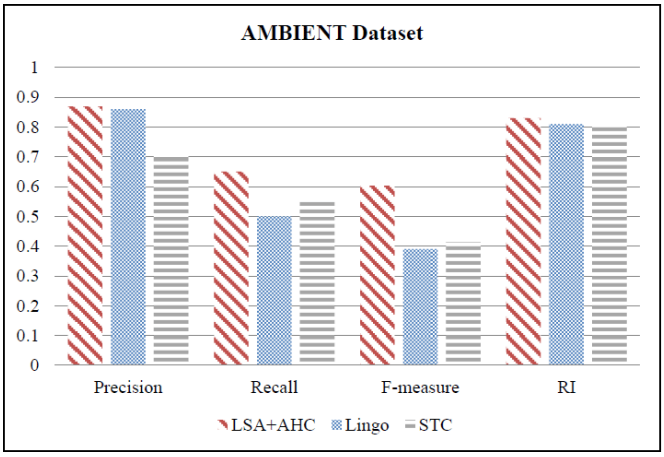
$$\text{RI} = (tp + tn)/(tp + tn + fp + fn) \quad \text{(식 8)}$$

$$F = \frac{2 \times \text{재현율} \times \text{정확도}}{\text{재현율} \times \text{정확도}} \quad \text{(식 9)}$$

위의 식에서 tp, tn, fp, fn은 각각 참 양성(true positive), 참 음성(true negative), 거짓 양성(false positive), 거짓 음성(false negative)를 나타낸다.

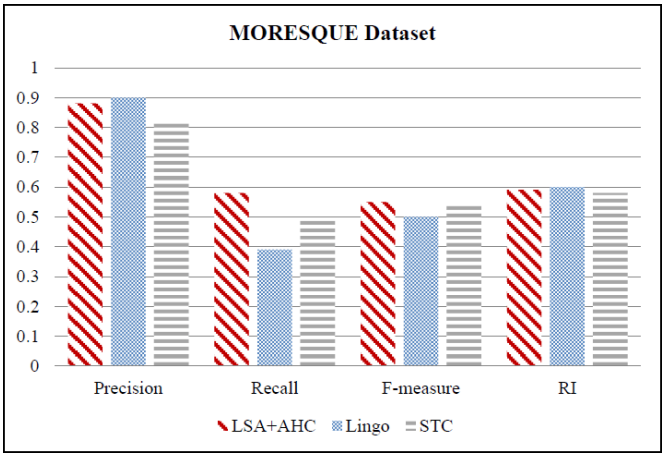
4.2. 결과

제안한 잠재 의미 분석을 이용한 응집 계층 군집화 방법의 효율성을 검증하기 위하여 기존의 군집화 방법인 서픽스 트리 군집화[1]와 링고(Lingo) [2] 방법과 비교한다. 그림 3 은 AMBIENT 데이터에 대한 다른 두 군집화 방법과 제안한 방법(LSA+AHC)을 앞 전에서 제시한 평가 기준에 따라 비교한 것이다.



(그림 3) AMBIENT 데이터에 대한 비교 결과

그림에서 보는 바와 같이 제안한 방법이 정확도, 재현율, F-측정(measure), 랜드 인덱스(RI) 측면에서 모두 우수함을 보인다. 반면, 그림 4 와 같이 MORESQUE 데이터에서는 재현율과 F-measure 에서만 더 좋은 결과를 보인다. 이는 MORESQUE 데이터가 AMBIENT 데이터보다 조금 더 복잡한 형태의 문장으로 구성되어 있기 때문이다. 이는 차후 자연어 처리 단계의 성능을 더 향상시킴으로써 해결 가능할 것으로 기대한다.



(그림 4) MORESQUE 데이터에 대한 비교 결과

따라서 본 논문에서 제안한 혼합형 방법이 기존의 방법들에 비해 재현율과 F-측정 측면에서 모두 우수함을 보인다. 그리고 본 논문에서는 문서 전체 데이터를 이용하지 않고 검색 엔진을 통한 검색 결과에서 많이 이용되는 스니핏을 이용하기 때문에 시간 복잡도가 높은 두 방법인 잠재 의미 분석 기법과 응집 계층 군집화 방법을 혼합할 수 있었다. 또한, 제안한 혼합형 방법이 더 질 높은 군집 성능을 나타냄을 실험을 통하여 보였다.

5. 결론 및 향후 연구

본 논문에서는 잠재 의미 분석 방법과 계층적 군집화 방법의 단점을 상호 보완하여 보다 효율적인 정보 검색을 위한 잠재 의미 분석 방법을 이용한 응집 계층 군집화 방법을 제안한다. 제안한 방법을 이용하여 기존의 방법들이 가지고 있는 백-오브-워드(bag-of-word) 문제, 단층적 군집 문제를 해결할 수 있다. 뿐만 아니라, 기존의 방법과 비교 실험을 통하여 군집의 질적 측면에서 볼 때 제안한 방법의 우수함을 보인다.

향후 연구로 실험에서 사용된 두 가지 알려진 데이터 외 다양한 데이터에 적용하여 더욱 광범위한 영역에서의 검증이 수행되어야 한다. 뿐만 아니라, 각 군집에 대한 라벨링(labeling) 연구가 추가적으로 수행되어 보다 더 사용자에게 정보 검색에서 있어서 효율성을 높인다면 지능형 정보 검색 시스템으로 거듭날 수 있을 것으로 기대한다.

사사

본 논문은 교육과학기술부의 재원으로 한국연구재단의 지원을 받아 수행된 BK21 플러스 사업의 연구 결과임 (No. T1300571)

참고문헌

[1] O. Zamir, O. Etzioni, Web document clustering: A feasibility demonstration, in: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, (ACM, 1998), pp. 46-54.
[2] S. Osiński, J. Stefanowski, D. Weiss, Lingo: Search results clustering algorithm based on singular value decomposition, in: Intelligent information processing and web mining, (Springer, 2004), pp. 359-368.
[3] S. Osiński, D. Weiss, Carrot2: Design of a flexible and efficient web information retrieval framework, in: Advances in Web Intelligence, (Springer, 2005), pp. 439-444.
[4] S. Koshman, A. Spink, B.J. Jansen, Web searching on the Vivisimo search engine, Journal of the American Society for Information Science and Technology, 57 (2006) 1875-1887.
[5] C. Carpineto, S. Osiński, G. Romano, D. Weiss, A survey of web clustering engines, ACM Computing Surveys (CSUR), 41 (2009) 17.
[6] S.C. Deerwester, S.T. Dumais, T.K. Landauer, G.W. Furnas, R.A. Harshman, Indexing by latent semantic analysis, JASIS, 41 (1990) 391-407.
[7] L. Zhao, J. Callan, Term necessity prediction, in: Proceedings of the 19th ACM international conference on Information and knowledge management, (ACM, 2010), pp. 259-268.
[8] G.W. Furnas, T.K. Landauer, L.M. Gomez, S.T. Dumais, The vocabulary problem in human-system communication, Communications of the ACM, 30 (1987) 964-971.
[9] P.C. Hansen, T. Sekii, H. Shibahashi, The modified truncated SVD method for regularization in general form, SIAM Journal on Scientific and Statistical Computing, 13 (1992) 1142-1150.
[10] L.N. Trefethen, D. Bau III, Numerical linear algebra, (Siam, 1997).
[11] D. Müllner, fastcluster: Fast hierarchical, agglomerative clustering routines for R and Python, J. Stat. Softw, 53 (2013) 1-18.
[12] D. Achlioptas, F. McSherry, Fast computation of low rank matrix approximations, in: Proceedings of the thirty-third annual ACM symposium on Theory of computing, (ACM, 2001), pp. 611-618.
[13] C. Carpineto, G. Romano, Ambient dataset, in, (2008).
[14] R. Navigli, G. Crisafulli, Inducing word senses to improve web search result clustering, in: Proceedings of the 2010 conference on empirical methods in natural language processing, (Association for Computational Linguistics, 2010), pp. 116-126.
[15] D.M. Powers, Evaluation: from precision, recall and F-measure to ROC, informedness, markedness & correlation, Journal of Machine Learning Technologies, 2 (2011) 37-63.