위키피디아에 기반한 단어 사이의 의미적 연결 관계 탐색

Discovering Semantic Relationships between Words by using Wikipedia

Ju-Hwang Kim, Min-sung Hong, O-Joun Lee, Jason J.Jung
Dept. of Computer Science and Engineering, Chung-Ang University

중앙대학교
CHUNG-ANG UNIVERSITY

Index

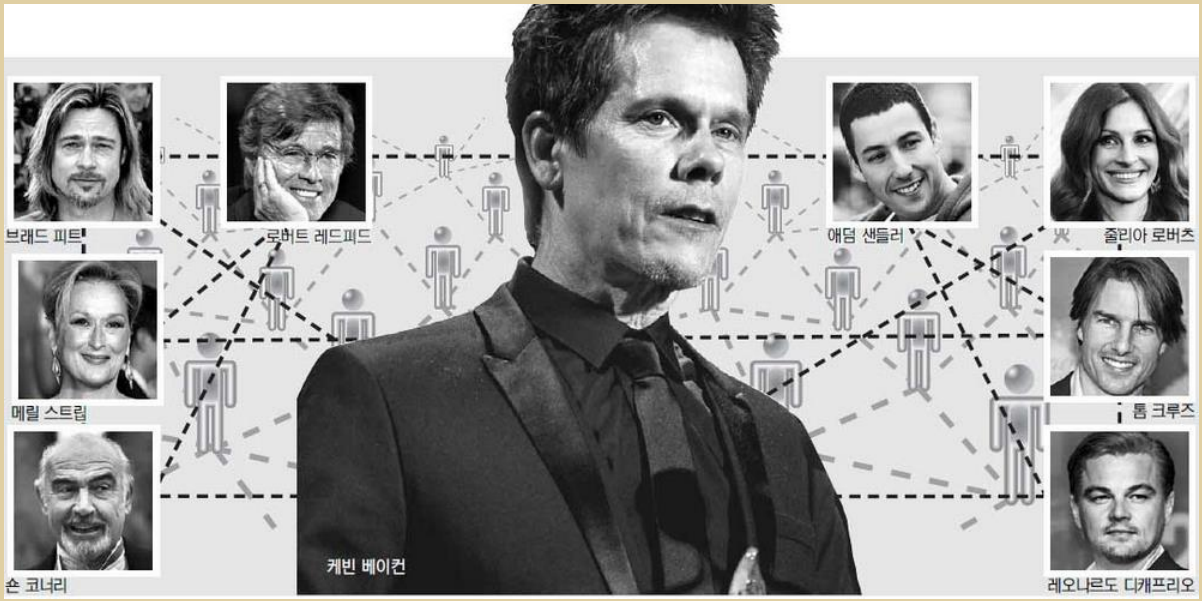
중앙대학교
CHUNG-ANG UNIVERSITY

01

02

03

1.1 케빈 베이컨의 여섯 다리 이론



- 어떤 배우와 케빈 베이컨까지 최단의 연결 고리를 만드는 놀이
- 메신저와 웹 사이트에서 많은 학술적인 연구 진행

중앙대학교 CHUNG-ANG UNIVERSITY

1.2 서론

01

02

03

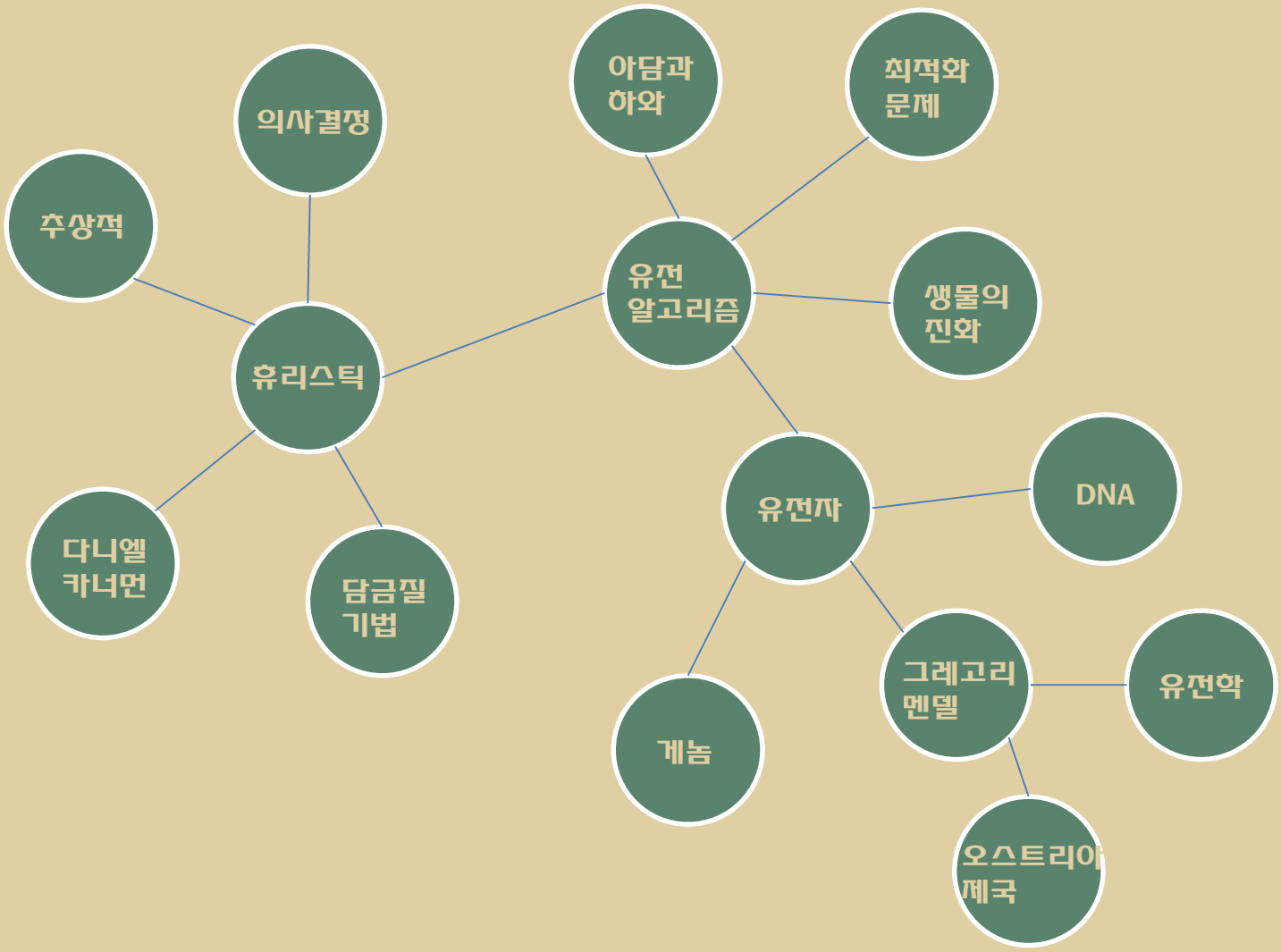
- 온톨로지나 워드넷 등 어휘 간 의미 유사도 기반의 다양한 연구들이 진행
- 단순히 어휘들의 유희어 집단을 형성하는 것에 그치지 않고 좀 더 폭 넓고 심도있는 의미 분석이 필요
- 위키피디아의 하이퍼링크 정보를 이용
- 단어 네트워크를 확장 및 이용

1.3 위키피디아 연결 정보 예

01

02

03



2. 그래프 구성 방식

01

02

03

- 위키피디아 API와 덤프 파일 이용
- 동적인 그래프 구성
- 정적인 그래프 구성

2.1 동적 그래프 구성 방식

```
<api batchcomplete="">
<limits links="500"/>
<query>
<pages>
<page _idx="-1" ns="0" title="아모스 트버스키" missing=""/>
<page _idx="-2" ns="0" title="어떻게 문제를 풀 것인가" missing=""/>
<page _idx="-3" ns="0" title="타부 서치" missing=""/>
<page _idx="-4" ns="0" title="평균회귀" missing=""/>
<page _idx="725782" pageid="725782" ns="0" title="다니엘 카너먼"/>
<page _idx="111968" pageid="111968" ns="0" title="담금질 기법"/>
<page _idx="1028335" pageid="1028335" ns="0" title="메타휴리스틱"/>
<page _idx="98821" pageid="98821" ns="0" title="발견적 평가방법"/>
<page _idx="181412" pageid="181412" ns="0" title="유전 알고리즘"/>
<page _idx="426491" pageid="426491" ns="0" title="의사결정"/>
<page _idx="117077" pageid="117077" ns="0" title="조합최적화"/>
<page _idx="828650" pageid="828650" ns="0" title="지역 최적해"/>
<page _idx="999670" pageid="999670" ns="0" title="추상적"/>
<page _idx="1234437" pageid="1234437" ns="4" title="위키백과:인용 오류 도움말"/>
</pages>
</query>
</api>
```

- REST 형식의 API 이용
- XML 형식을 얻는 과정을 재귀적으로 호출

2.2 동적 그래프 알고리즘

```
G : words of graph
L : the set of linked word
Q : queue
a, b : input words
c : depth of graph
w : selected word

Q <- a
While c < proper upper bound of c
  w <- Q
  If w == b
    path <- find path of graph from a to b
    return path
  End if
  L <- linked words of w using wiki API
  For i = 1 to i = number of L
    w_i ∈ L
    If  w_i □is not visited
      G <-  w_i
      Q <-  w_i
    End If
  End For
End While
```

- BFS를 이용한 탐색
- 일정한 단계수를 지정하여 그래프의 크기를 제어

2.3 정적 그래프

01

02

03

- 위키피디아의 덤프 파일을 이용
- 그래프 구성을 데이터베이스 화
- 그래프의 최단 경로 문제와 동일
- 탐색으로는 데이크스트라 알고리즘을 이용

2.4 유사도 계산

01

02

03

- 연결된 간선의 개수에 따라 유사도 계산이 가능
- 유사도를 간선의 합으로 표현
- 탐색을 통해 빠르게 유사도 계산이 가능

3. 결론 및 향후 연구 계획

01

02

03

- 위키피디아 문서내 하이퍼링크를 이용
- 간단하고 넓은 단어의 도메인을 포괄할 수 있다는 장점이 있음
- 향후 단어의 연결 고리를 이용한 새로운 연구에 도전
- 정밀한 유사도 계산 방법에 대한 연구

Thank You