Text-Mining on the Web

- 기반기술과 전망 -

임 일

연세대학교 경영대학

Agenda

- Text-mining 소개
- 기반 기술
- 응용분야
- 전망 및 연구분야

Text-Mining이란?

- Text (contents)를 분석하는 것
- 인터넷의 문제 – information overload
- 해결 방법
  - 사람으로부터의 정보 활용해서 필터링 – 평가 (review), Collaborative filtering (CF)
  - 텍스트를 분석 – 자연어처리, Text-Mining

Text-Mining 기술의 과제

- 텍스트의 내용을 컴퓨터가 어떤 형태로든 이해하여야 함
- 이상적인 방법: 자연어 처리
  - 컴퓨터가 텍스트의 내용을 완벽하게 이해
  - 기술적 한계가 많음
- 현실적인 방법: Text-Mining
  - 단어의 뜻은 모르더라도 각 단어의 종류, 중요도, 관계를 분석
  - 가능하며 상당히 유용한 정보를 제공할 수 있다
- 인간 언어의 복잡성으로 인해 각 단어의 종류, 중요도, 관계를 분석하는 것도 쉽지는 않다
- 방대한 양의 텍스트를 긁어오는 효율적인 Crawler의 개발

Text-Mining 기술 1 : 단어의 추출

- 형태소 분석기 – 텍스트에서 각 단어의 종류와 원형 등을 추출해 내는 프로그램
  - 연세대학교 e-lab (http://e-lab.yonsei.ac.kr/lexical/test.jsp)
  - Zalab (http://lab.zagia.com/)
  - 한국과학기술정보연구원(KISTI) 연구 (http://www.kristalinfo.com/)
  - 서울대학교 지능형 데이터베이스 연구실 (http://ids.snu.ac.kr/wiki/Implementing_typo_and_spacing_tolerant_korean_morpheme_analyzer)
- 영어권에서는 대부분 Noun extractor 활용

Text-Mining 기술 2 : 추출된 단어 중 핵심어구 판별 (Keyphrase Identification)

- 추출된 단어 중 핵심 단어 혹은 어구 (Keyword / Keyphrase)를 판별
- 중요한 단어는 두 가지 조건을 충족시켜야 한다
  - 그 문서에는 자주 등장
  - 다른 문서에는 적게 등장
- 이것을 이용해서 단어의 중요도를 평가하는 기법이 tf.idf (term-frequency / inverse document frequency)

tf.idf (Term Frequency – Inverse Document Frequency)

- $IDF = ln(N/d_k)+1 = ln(N) – ln(d_k)+ 1$
- $w_{ik} = TF * IDF = f_{ik} * (ln(N) - ln(d_k) + 1)$

  N: the number of documents in the document collection,

  d_k: the number of documents containing the term *k*,

  f_{ik}: the absolute frequency of term *k* in document *i,* and

  w_{ik}: the weight of term *k* in document *i*
- Example of final output

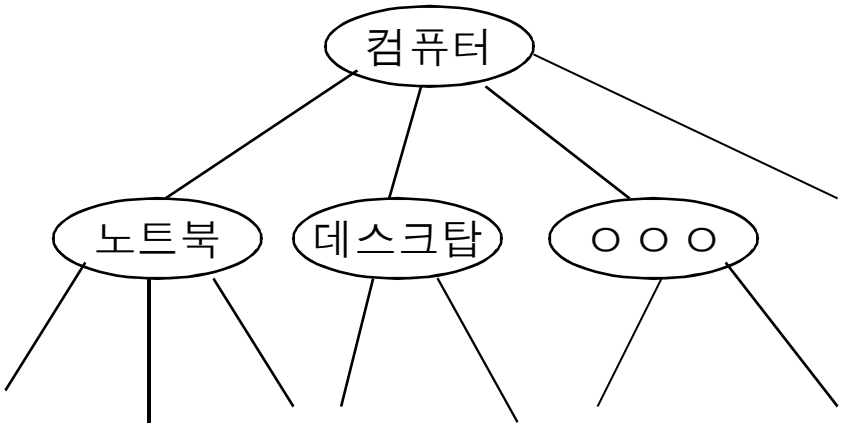
Document 1	. . .	**Document n**
computer – 5.321 html – 2.342 xml – 2.112 web – 1.980 . . .		html – 4.321 computer – 3.231 java – 3.123 web – 2.311 . . .

Text-Mining 기술 3 : 추출된 핵심어 구의 활용

- 내용기반 필터링 (content-based filtering)에 활용
  - 문서간 유사도를 추출된 keyword score를 이용해서 계산한다.
  - 예) AIS 문서 검색 시스템 (by Brook Wu) (http://highlight.njit.edu/ais)
- Keyword extension
  - 한 키워드와 같이 사용된 키워드를 검색에 같이 활용

추출된 중요단어의 활용 - 개념 지도

- 추출된 키워드의 관계를 분석해서 한 도메인에서의 concept map을 구축
- 구축방법의 예 – POCA (probability of co-occurrence analysis) (Wu, 2001)
- 구축된 concept map의 예
  - P(노트북|컴퓨터) = 0.4
  - P(데스크탑 |컴퓨터) = 0.3
  - P(컴퓨터|노트북) = 0.8
  - P(컴퓨터|데스크탑) = 0.9
  - P(ㅇㅇㅇ|ㅇㅇㅇ) = 0.2
  - ⋮



Concept Map의 활용

- Concept map을 하나의 ontology로 활용
  - Concept map 상에서 가까운 키워드를 사용한 keyword extension
- Concept map을 이용해서 검색 결과를 분류
  - 검색된 문서의 keyword와 concept map을 비교해서 해당 문서를 적절한 카테고리로 분류
  - 예: Highlight 메타 검색 엔진 (by Brook Wu) – 여러 검색엔진에서 받은 검색결과를 concept map을 이용해서 분류해 준다 (Demo: http://highlight.njit.edu)

Text-Mining의 응용

- 검색엔진 정확도 향상 (Keyword extension, 문서간 유사도 사용)
- 추천시스템
  - 복합추천시스템 (http://gre.njit.edu)
- 온라인 시장조사
  - Umbria Communications (http://www.umbrialistens.com/)
  - 실시간 시장조사가 가능

Text-Mining의 전망과 과제

- 완전한 자연어 처리가 궁극적인 목표
  - 정확도, 처리속도 향상이 필요
  - 단어의 의미 (최소한 동의어/반의어에 대한) 에 대한 처리 기술이 필요 → 정확하고 방대한 전자사전이 필수
- 다른 기술과의 결합
  - 추천기술 (Collaborative filtering), 검색 엔진, 상황인식 기술 등
- 기업에서의 활용방법 확립
  - 현재 Text-Mining을 마케팅이나 다른 분야에 어떻게 활용할 것인가에 대한 확립된 practice가 없다

Text-Mining 관련 연구 과제

- 기술적인 분야
  - 좀 더 정확한 형태소 분석기, 핵심어구 판별기, 응용방법 등
- 활용 분야
  - Text-Mining이 더 적합한 분야는 무엇인가?
  - Text-Mining으로 수집된 정보의 효과적인 활용 방법은 무엇인가?
  - Text-Mining 결과를 사람들이 어떻게 받아 들이는가? 혹은 더 잘 받아들이도록 하는 방법은?
  - Text-Mining을 활용한 새로운 서비스/비즈니스 모델에는 어떤 것이 있으며 어떤 것이 효과적인가?