STS2008
Search Technology Summit

www.stskorea.org

Semantic Search:
자연어처리와 클러스터링기술의 활용

㈜코난테크놀로지
장 정 호

2008. 9. 2

STS
Search Technology Summit

STS2008
Search Technology Summit

목 차

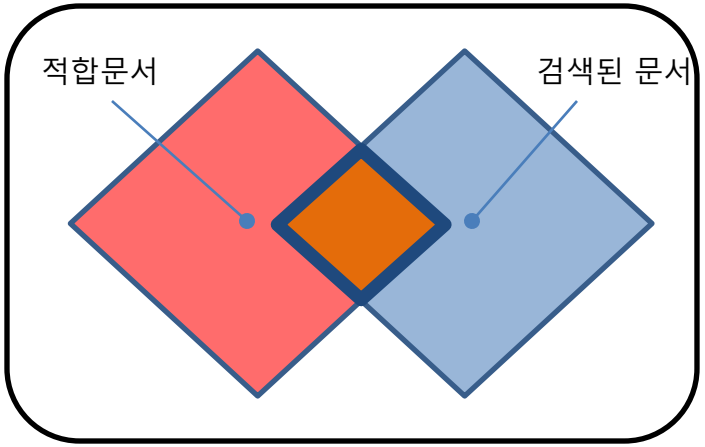
- Semantic Search 개요
- 자연언어처리 기술의 활용
- Clustered Search

Semantic Search

"Methods of searching web documents beyond the syntactic level of matching keywords…
Semantic search - defined as an IR with the capabilities to understand the user's intent and the web's content at a much deeper, conceptual level."

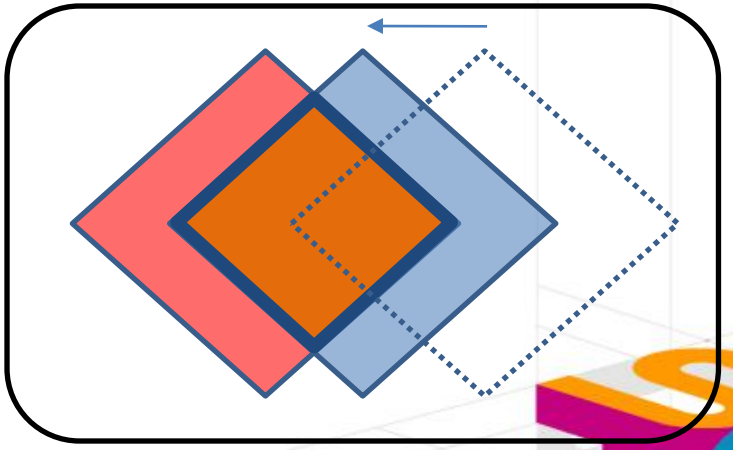
- from "Semantic search arrives at the Web" by Peter Mika, Yahoo!

키워드 매칭 기반 검색



Semantic search

질의 의미를 파악하고
의미 수준에서 적합 문서 검색



검색결과의 정확도/적합성 향상

Semantics in IR: An Example

Semantic

Syntactic

이순신의 출생지는?


출생지


이순신(인물, ) 서울 건천동 (지명, )

태어나다

이순신 건천동

서울

이순신, 서울, 건천동, 태어나다

이순, 순신, 신은, 은서, 서울, 울건, 건천, 천동, 동에, 에서, …

"이순신은 서울의 건천동에서 태어났다"

모네가 태어난 곳은?


출생지


끌로드 모네(인물, ) 프랑스 파리(지명, )

출생하다

끌로드 모네 파리

프랑스

끌로드, 모네, 프랑스, 파리, 출생하다

끌로, 로드, 드모, 모네, 네는, 는프, 프랑, 랑스, 스의, 의파, 파리,

"끌로드 모네는 프랑스의 파리에서 출생했다."

검색 단계에 따른 의미 정보 응용 예 (semantic + IR)

STAGE	FUNCTIONALITY
Query Construction	• **입력 키워드의 의미모호성 해소** 파리 ➔ do you mean...?   • **키워드 추천** 파리(Paris) ➔ 파리교통정보, 파리관광지, 파리여행상품, ...
Search Process	•**질의문 의미 분석/확장** 단어수준: 유의어: 비행기, 항공기, 구체화: 자동차 가격 (현대차, 기아차, GM대우차..) 구/문장수준: 정규화: 이순신이 태어난 곳은? (이순신 출생지) •**의미에 기반한 매칭 (semantic matching)** : 색인 단계에서의 의미 표현을 고려.
Presentation of Search Result	•**검색 결과의 조직화** 주제/속성별로 분류화된 결과 제시(트리, 맵) •**사용자 피드백** 질의문 정제/확장 관련 주제/키워드/리소스 추천

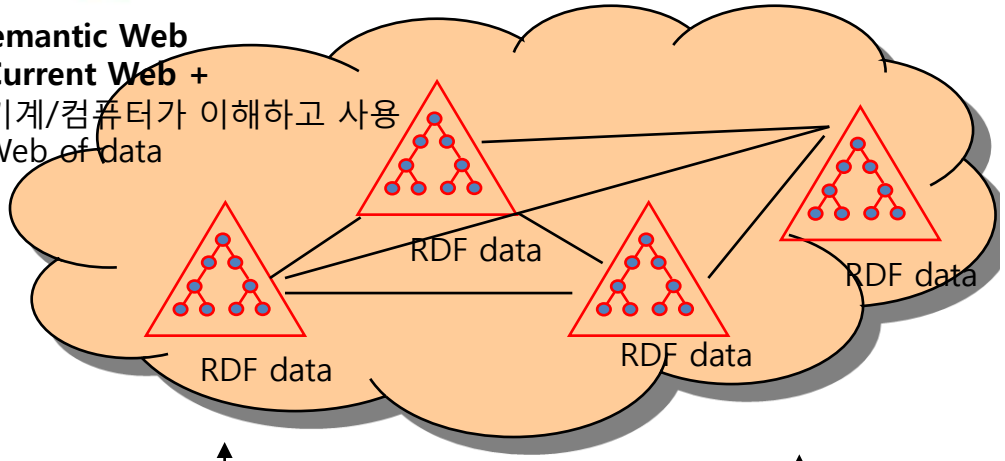
Semantic Sources for Semantic Search

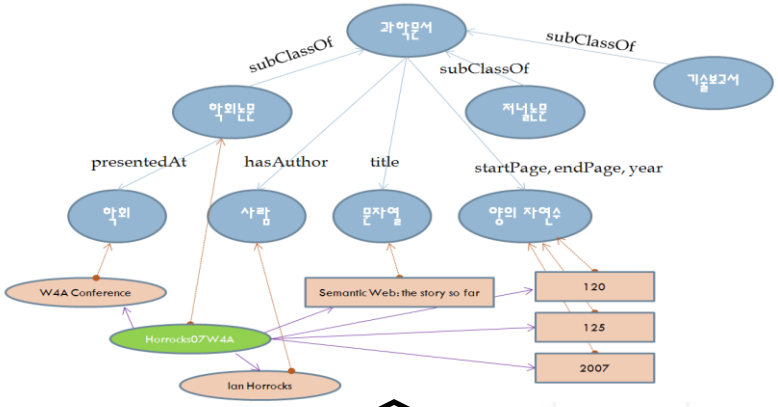
- **Explicit Semantics**
  - Top-down.
  - Described in a structured way from the scratch or referring to existing ones.
  - Semantic Web technology.

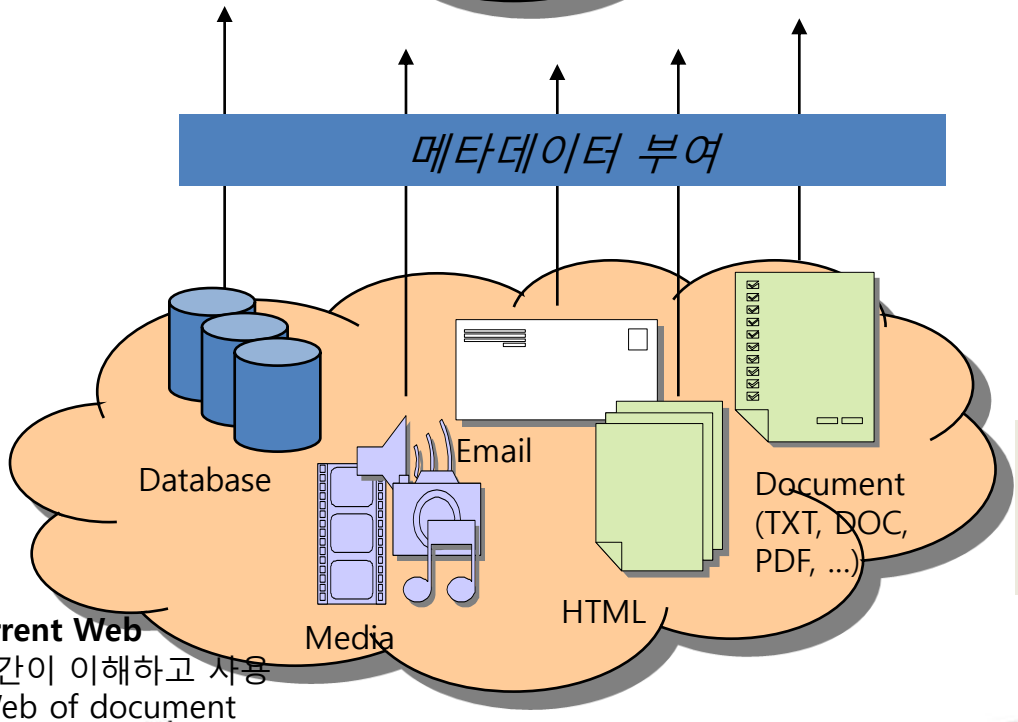
- **Exploitation of Implicit Semantics**
  - Bottom-up.
  - semantic structures are extracted/identified from unstructured texts.
  - Techniques including information extraction, natural language processing, and classification (categorization, clustering) can be used.

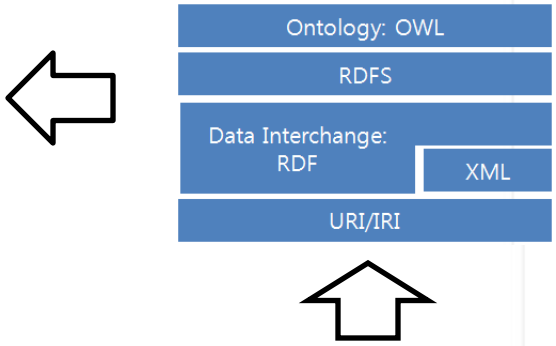
Semantic Search: 자연어처리와 클러스터링기술의 활용

Semantic Web
-Current Web +
-기계/컴퓨터가 이해하고 사용
-Web of data







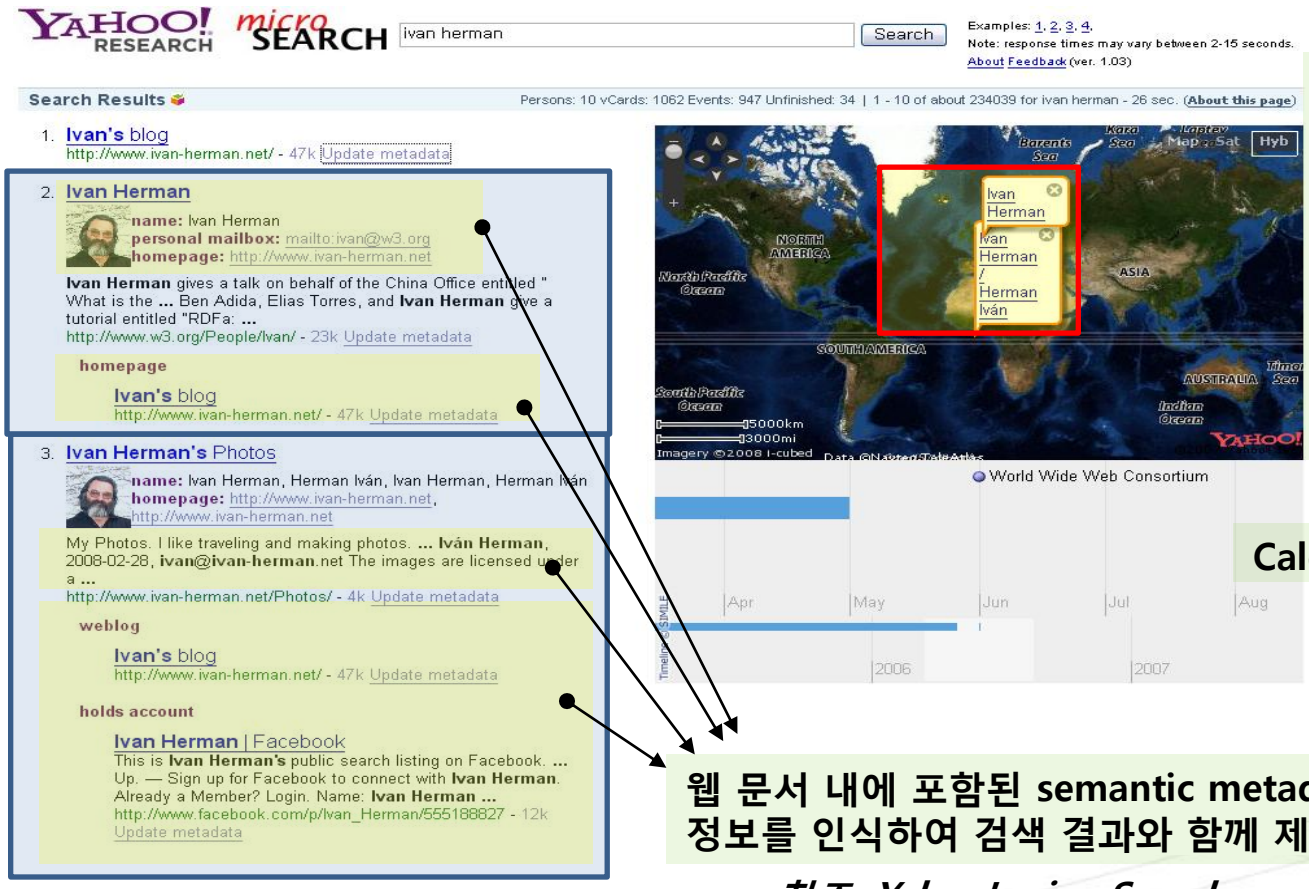


Current Web
-인간이 이해하고 사용
- Web of document

Ian Horrocks, "Semantic Web: the story so far", In *Proceedings of the 2007 international cross-disciplinary conference on Web accessibility* (*W4A*), pages 120-125 , 2007.

Search Supporting Explicit Semantics: An Example

웹 문서 내에 삽입된 의미 정보(eRDF, RDFa, microformat 구문에 의한 metadata)의 추출하여 확장된 검색 제공.



웹 문서 내에 포함된 semantic metadata 정보를 인식하여 검색 결과와 함께 제시.

참조: Yahoo! microSearch

DBpedia
http://dbpedia.org/

Wikipedia article들을 대상으로 구조화된 의미정보를 추출하여 웹 상에서 접근할 수 있도록 하기 위한 프로젝트

Wikipedia 문서

Taipei 101

From Wikipedia, the free encyclopedia

description

Taipei 101 (traditional Chinese: 臺北101 or 台北101; simplified Chinese: 台北101; pinyin: *Táiběi Yīlíngyī*; Wade-Giles: *T'ai-pei I-ling-i*; POJ: Tai-pak It-leng-it) is a 101-floor landmark skyscraper located in Xinyi District, Taipei, Taiwan. The building, designed by C.Y. Lee & Partners[2] and constructed primarily by KTRT Joint Venture[3] and Samsung Engineering & Construction, is the world's tallest completed skyscraper according to the CTBUH[4] - the arbiter of tall building height. Taipei 101 received the Emporis Skyscraper Award in 2004. It has been hailed as one of the Seven New Wonders of the World (*Newsweek* magazine, 2006) and Seven Wonders of Engineering (Discovery Channel, 2005).[5]

The building stands as an icon of Taipei and Taiwan as a whole. The building contains 101 floors above ground and 5 floors underground. Its postmodern style combines tradition and modernity in ways that appear simultaneously Asian and international. Its safety features enable it to withstand typhoons and earthquakes. A multi-level shopping mall adjoining the tower houses hundreds of fashionable stores, restaurants and clubs. Fireworks launched from Taipei 101 feature prominently in international New Year's Eve broadcasts, and the landmark appears frequently in films, television shows, print publications, anime media, games, and other elements of popular culture.

The name of the tower reflects its location in Taipei's international business district (101 mailing code) as well as its floor count. (See also *"Symbolism"* below.) The number is pronounced in English simply as *One Oh One* and in Mandarin and other local languages by the equivalent.

Taipei 101 is owned by the Taipei Financial Center Corporation and managed by the International division of Urban Retail Properties Corporation based in Chicago. The name originally planned for the

image

Taipei 101 has been the world's tallest building since 2004.*	
Preceded by	Petronas Twin Towers
Information	
Location	Xinyi District, Taipei, Taiwan
Status	Complete
Constructed	1999-2004
Height	
Antenna/Spire	509.2 m (1,670.60 ft)
Roof	449.2 m (1,473.75 ft)
Top floor	439.2 m (1,440.94 ft)
Technical details	
Floor count	101
Floor area	412,500 m² (4,440,100 sq ft)
Elevator count	61, including double-deck shuttles and 2 high speed observatory elevators)
Cost	NT$58 billion (US$1.76 billion)[1]
Companies	
Architect	C.Y. Lee & partners
Contractor	KTRT Joint Venture, Samsung Engineering & Construction
Owner	Taipei Financial Center Corp.
Management	Urban Retail Properties Co.

infobox

검색: Tallest Buildings

Taipei 101	101	449.2m
Empire State Building	102	381 m
Sears Tower	108	442.3 m
World Trade Center	110	417.0 m
Chrysler Building	77	282.0 m

* http://wikipedia.3ba.se/ 참조

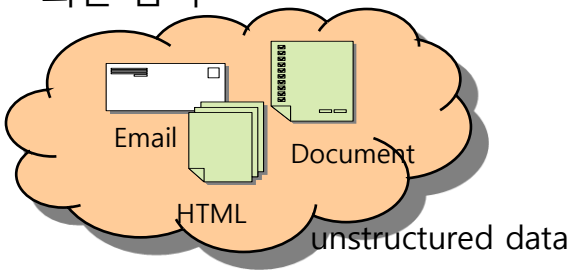
의미 정보 추출

Taipei 101

- *height_roof* → 449.2m (1,473.75ft)
- *floor_count* → 101
- *constructed_year* → 1999-2004
- • • •

Exploiting Implicit Semantics

질의 키워드의 단순 매칭에 의한 검색



Discover/extract relevant structure

질의 의도(의미)를 파악하고 의미 차원에서 부합하는 문서나 정보를 제공

정보의 범주화

자연어처리

키워드 수준

구/문장 수준

언어자원과 자연어처리 기술을 활용하여 질의/웹 문서의 단어/문장들에 대해 의미 분석을 수행하고 질의 의도에 적합한 문서나 정보 제공

클러스터링

검색 결과를 구축된 분류 체계를 통하거나 자동적으로 범주화 /구체화

키워드 context 제시

언어자원 | 자연어처리 | 인공지능 | 기계학습

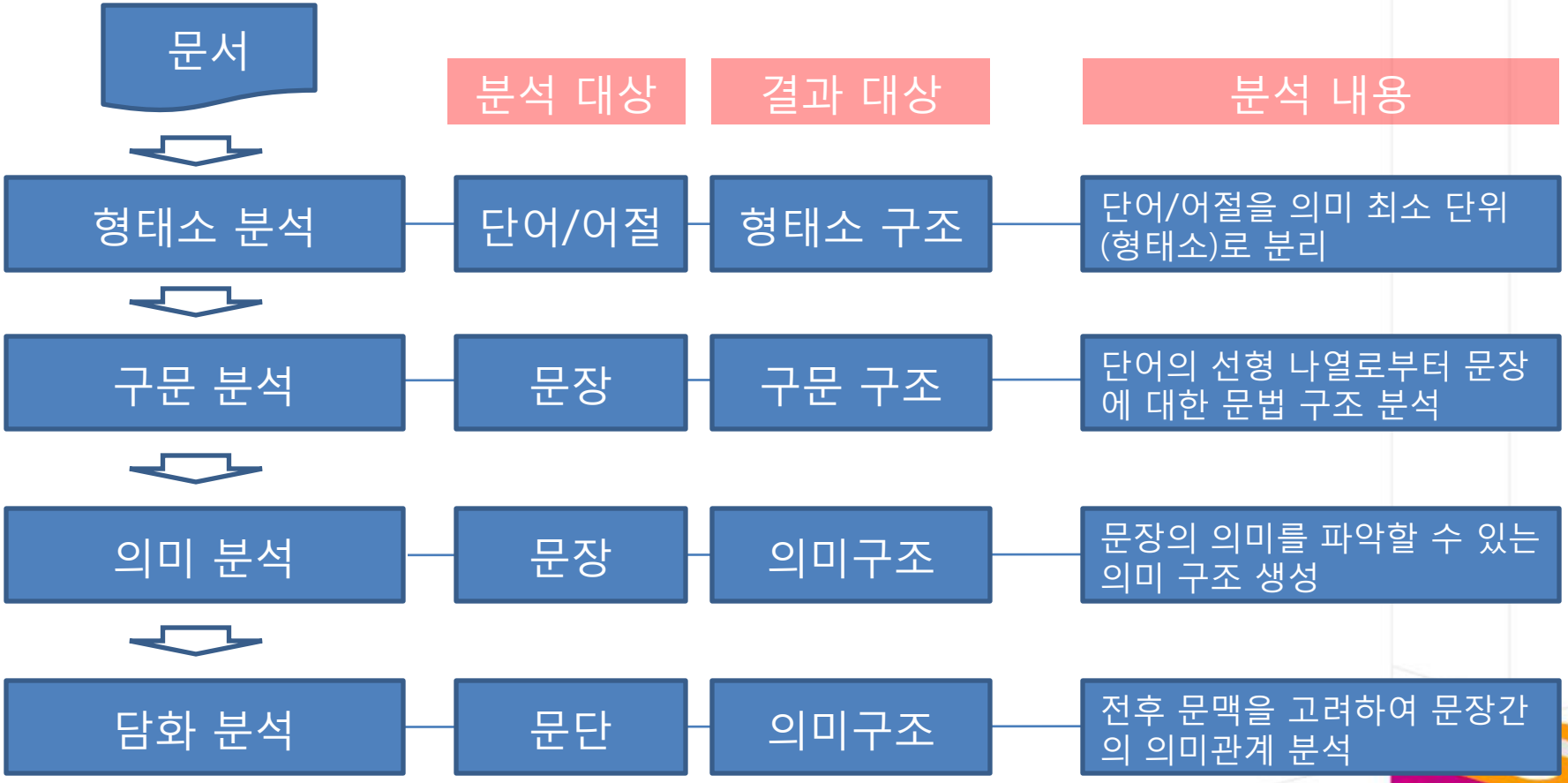
STS2008
Search Technology Summit

www.stskorea.org

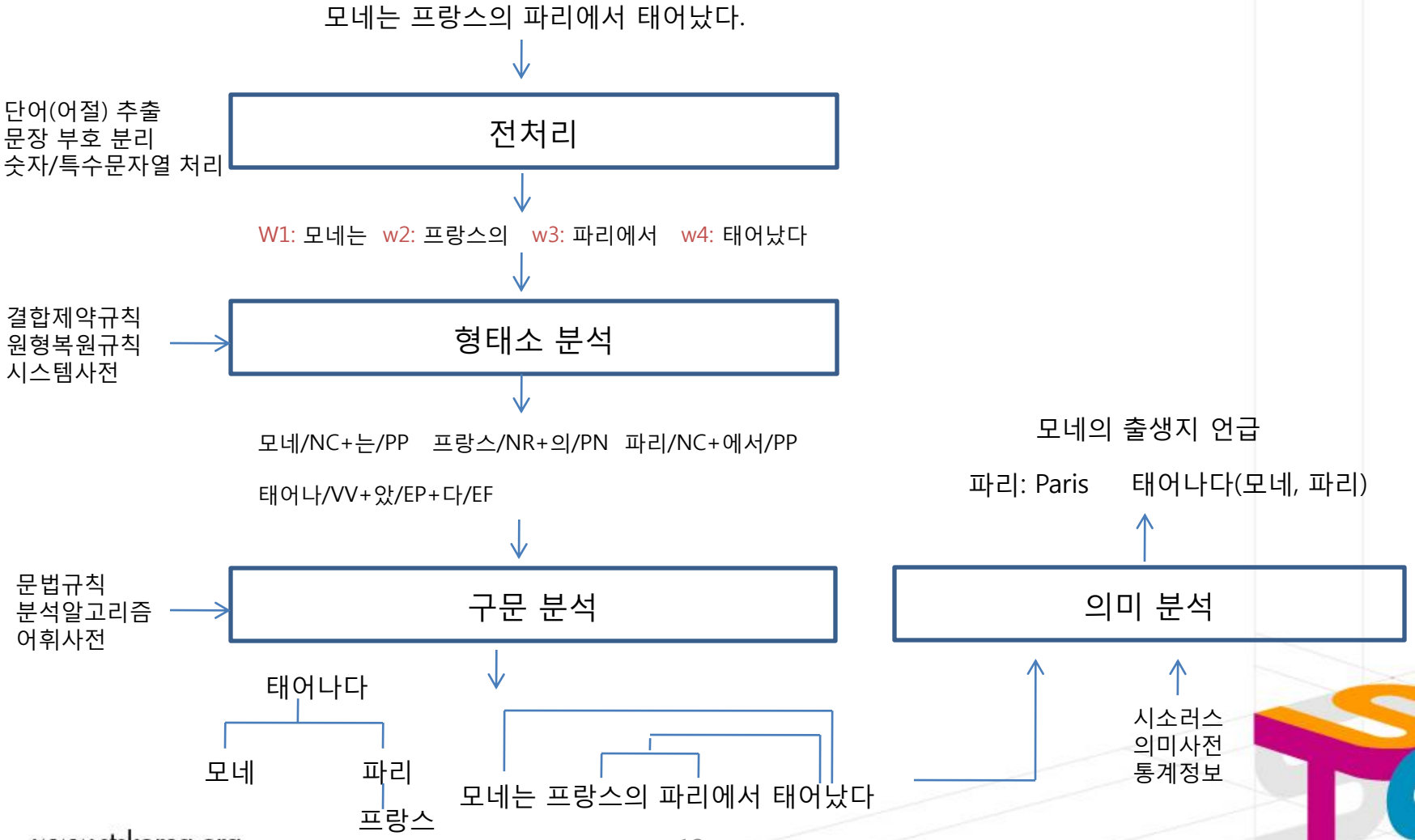
NLP 기반 semantic search

STS
Search Technology Summit

자연어 분석 과정


문서
분석 대상
결과 대상
분석 내용
형태소 분석
단어/어절
형태소 구조
단어/어절을 의미 최소 단위 (형태소)로 분리
구문 분석
문장
구문 구조
단어의 선형 나열로부터 문장에 대한 문법 구조 분석
의미 분석
문장
의미구조
문장의 의미를 파악할 수 있는 의미 구조 생성
담화 분석
문단
의미구조
전후 문맥을 고려하여 문장간의 의미관계 분석

NLP에 의한 문장 분석 예



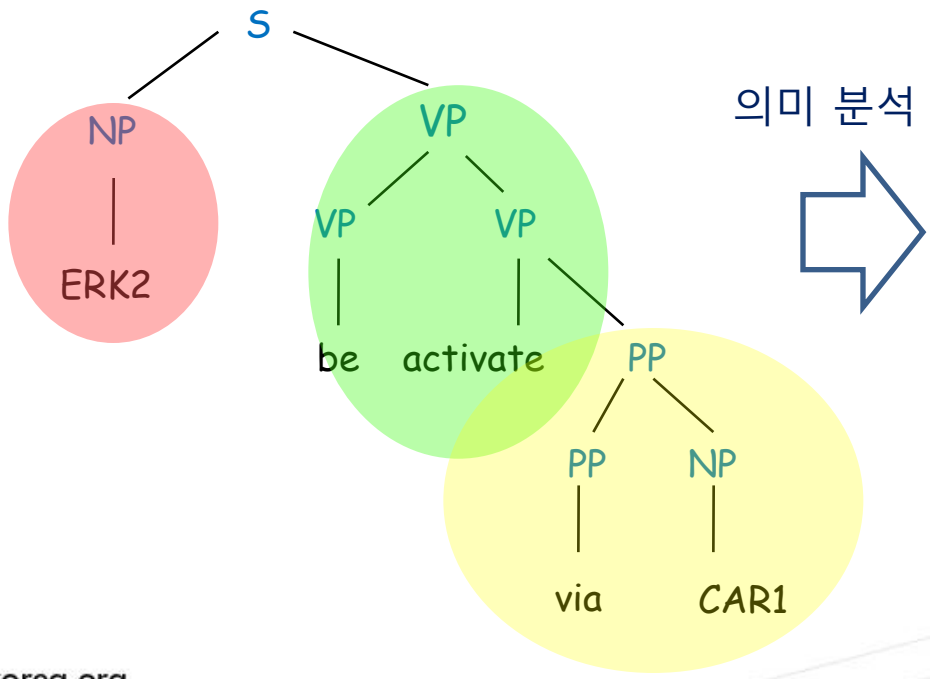
언어자원과 NLP기반 분석에 기초한 semantic search

"... ERK2 is activated via CAR1 ... "

형태소 분석

... ERK2/NN is/VBZ activated/VBN via/IN CAR1/NN...

구문 분석



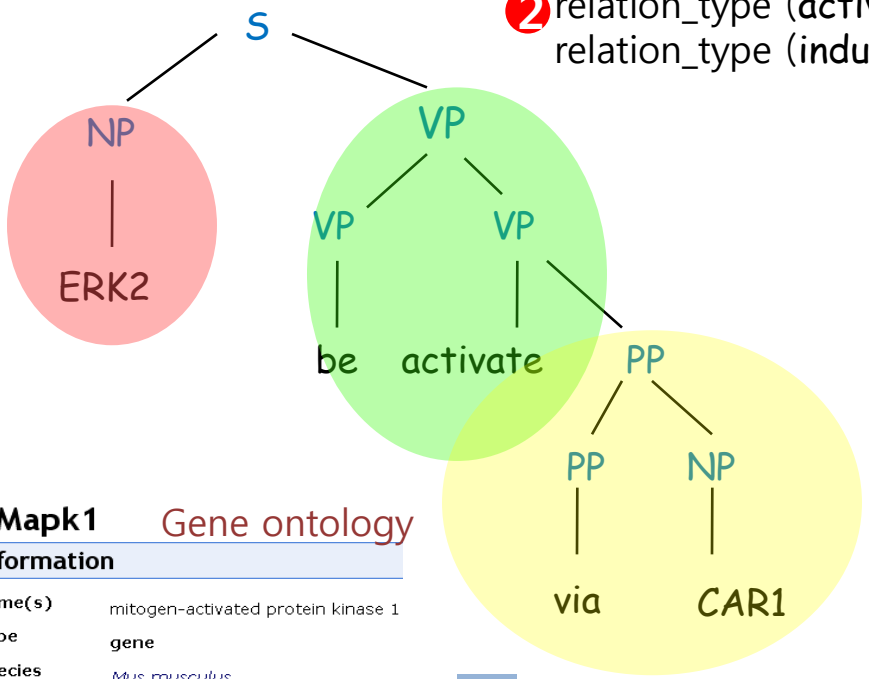
의미 분석

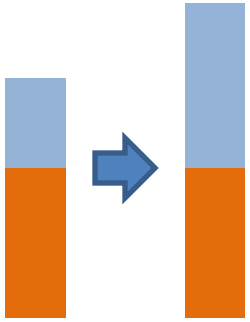
ERK2: gene
CAR1: gene
rel_activation(CAR1, ERK2)

Query: What induces MAPK1?

Word ontology

❷ relation_type (**activate**) = '*activation*'
relation_type (**induce**) = '*activation*'



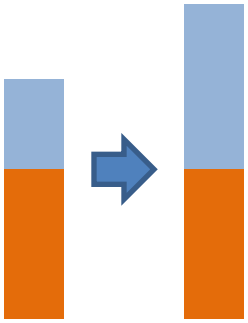


recall 향상

❶ **Mapk1** Gene ontology

Information

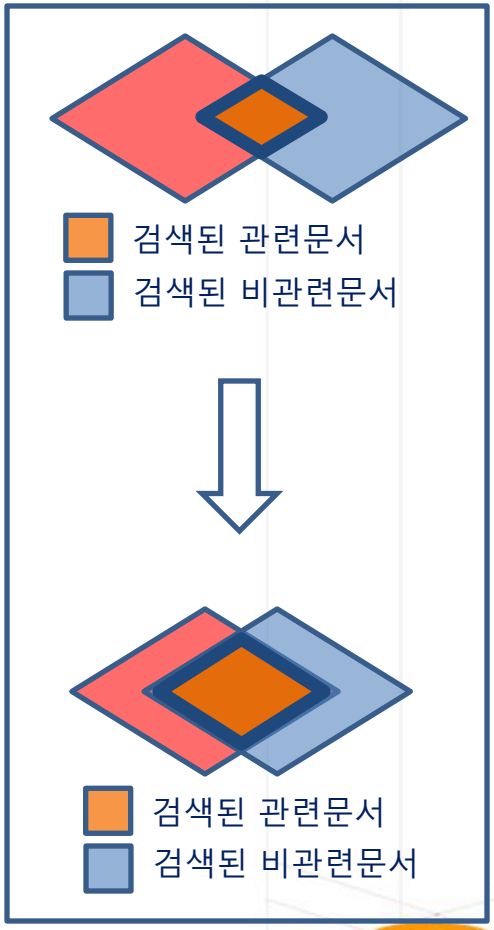
Name(s)	mitogen-activated protein kinase 1
Type	gene
Species	*Mus musculus*
Synonyms	Erk2 IPI00119663 MAPK2 p42mapk Prkm1



recall 향상

❸ rel_activation(*CAR1, ERK2*)

precision 향상



NLP 기반 Semantic Search 엔진 사례

Cognition

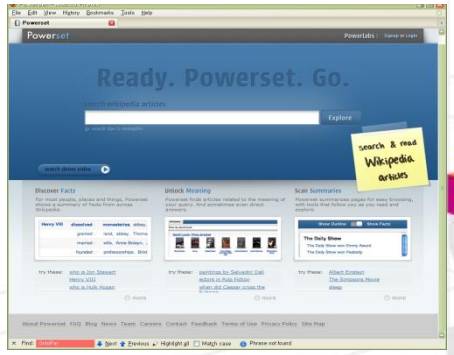
"… technology that *understands word and phrase meanings within context* in mod ern computer applications…" (from htt p://www.cognition.com)

Hakia

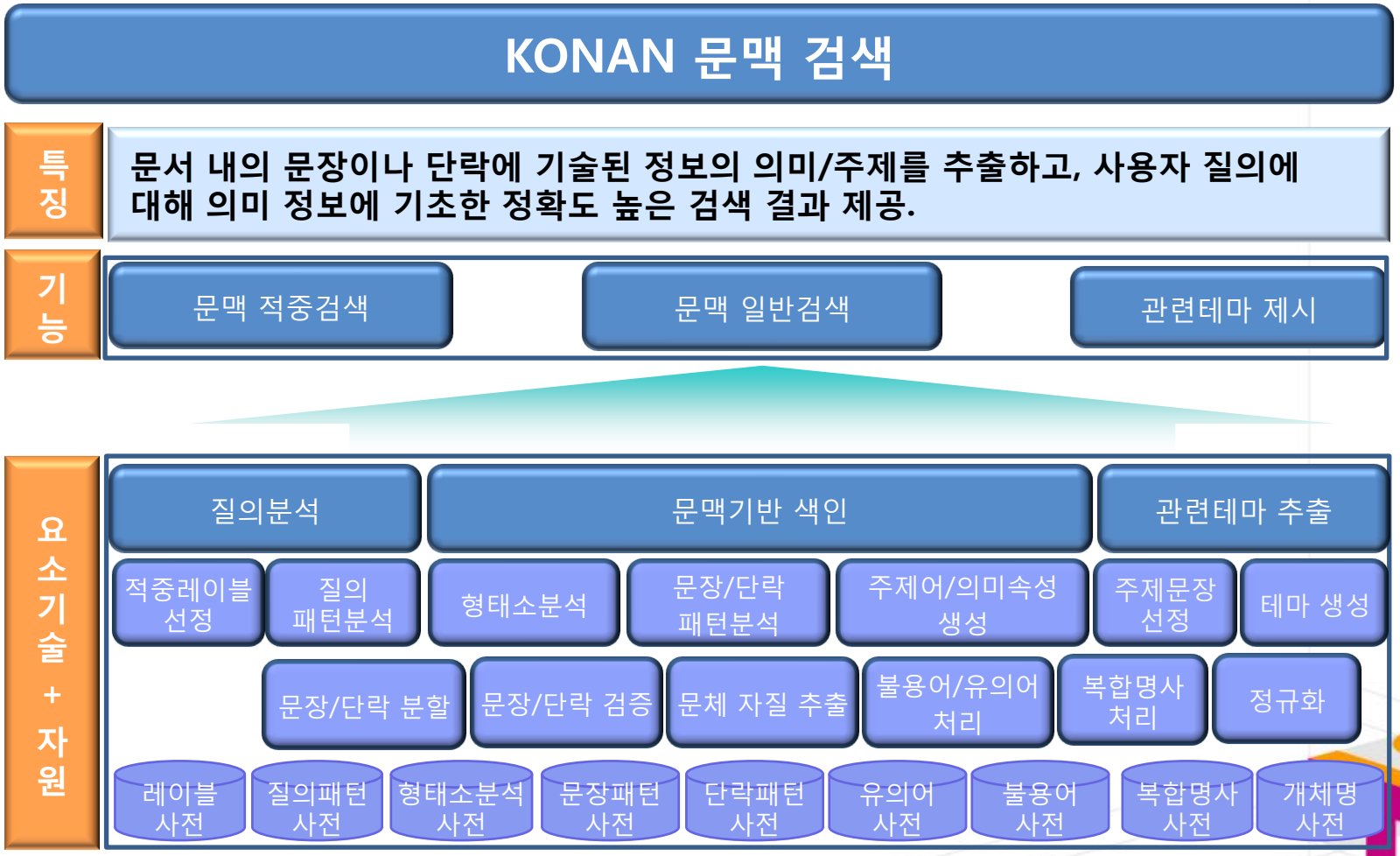
"… a semantic search engine that brings r elevant results based on *concept matc h rather than keyword match* or popula rity ranking." (from http://www.hakia.c om)

PowerSet

"… applying its *natural language proce ssing* to search, aiming to improve the way we find information by *unl ocking the meaning* encoded in ordi nary human language." (from http:/ /www.powerset.com)



KONAN 문맥 검색™


KONAN 문맥 검색
특징
문서 내의 문장이나 단락에 기술된 정보의 의미/주제를 추출하고, 사용자 질의에 대해 의미 정보에 기초한 정확도 높은 검색 결과 제공.
기능
문맥 적중검색
문맥 일반검색
관련테마 제시
요소기술 + 자원
질의분석
문맥기반 색인
관련테마 추출
적중레이블 선정
질의 패턴분석
형태소분석
문장/단락 패턴분석
주제어/의미속성 생성
주제문장 선정
테마 생성
문장/단락 분할
문장/단락 검증
문체 자질 추출
불용어/유의어 처리
복합명사 처리
정규화
레이블 사전
질의패턴 사전
형태소분석 사전
문장패턴 사전
단락패턴 사전
유의어 사전
불용어 사전
복합명사 사전
개체명 사전

문맥 검색™

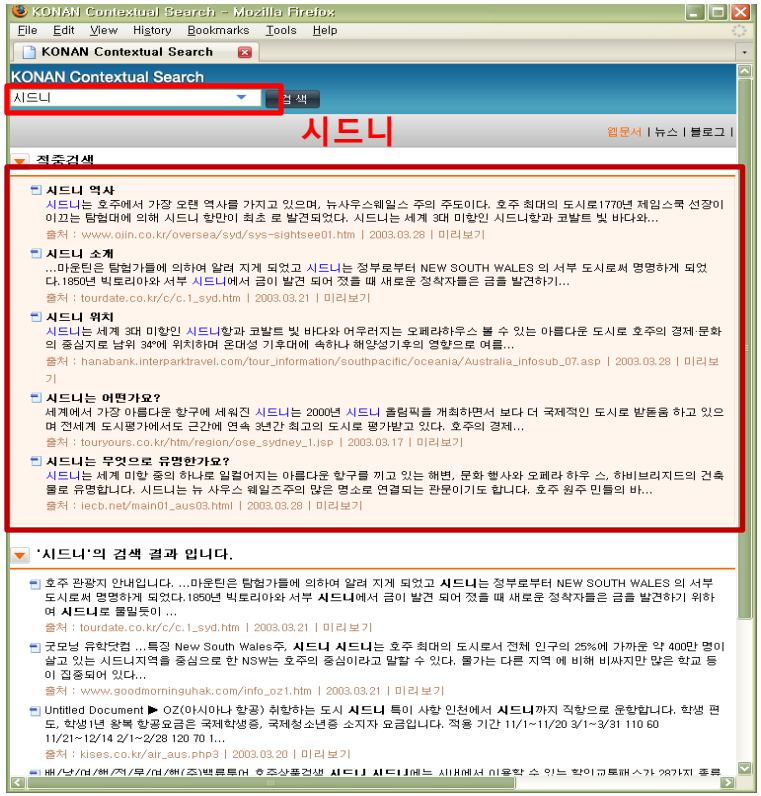
- **기존 검색과의 비교**

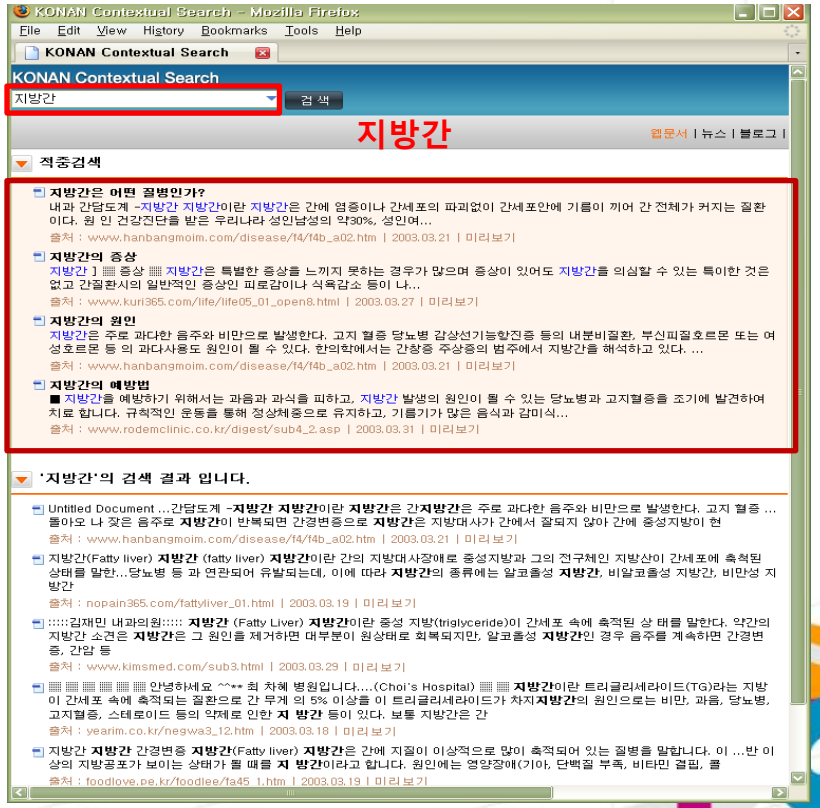
	N-GRAM 검색		형태소분석 검색		문맥 검색
Primitive	자소	⇒	품사	⇒	문맥
Key Extraction	Context-Free	⇒	Context-Free	⇒	Context-Sensitive
Key Form	String	⇒	Word	⇒	Phrase
User Intention	No	⇒	No	⇒	Yes
Semantic Search	No	⇒	No	⇒	Yes
Relevance (Quality)	Poor	⇒	Good	⇒	Excellent
Ranking	통계기반	⇒	통계기반	⇒	의미기반(적중검색)
적용 분야 예	책, 영화 제목	⇒	요약문	⇒	본문

문맥 검색™: 검색 예 (1)

검색 의도가 구체화되지 않은 단일 키워드 검색:
문서상에서 질의키워드가 사용된 문맥에 기초하여 의미 범주별로 클러스터링된 적합 문서 제시

'시드니'에 대한 범주별 정보

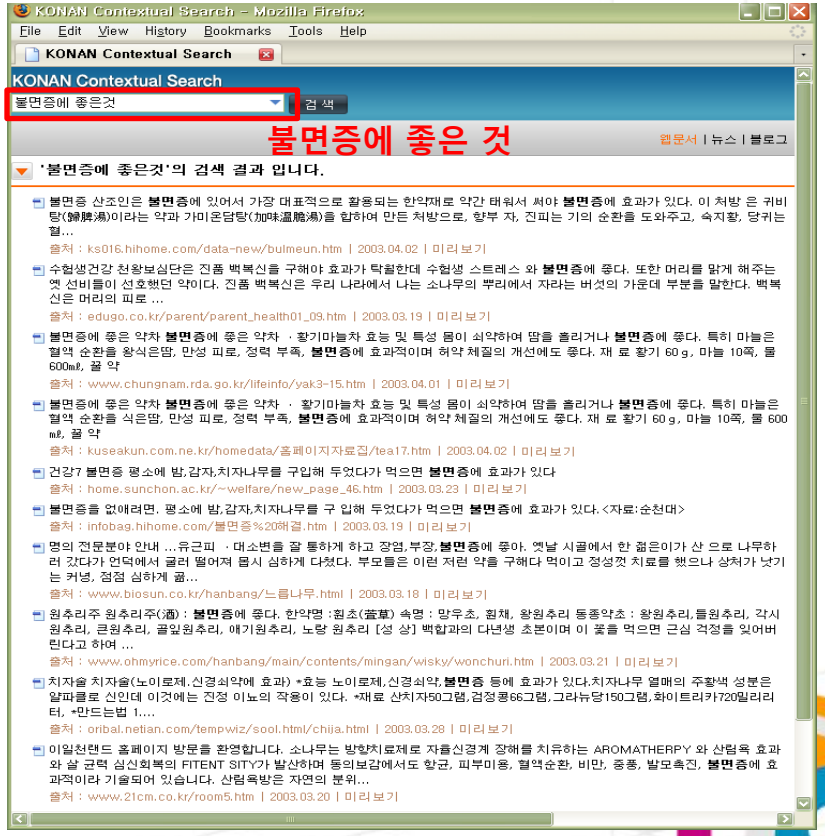




문맥 검색™: 검색 예(2)

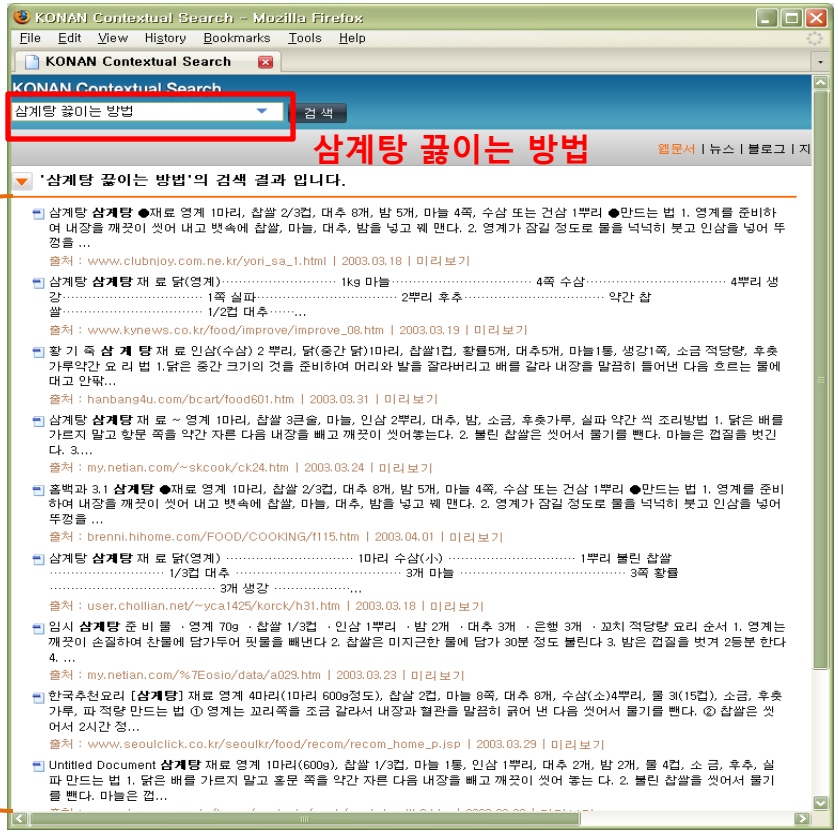
구체적 질의 의도에 대해 의미적으로 적합한 문장이 서술된 문서 검색

발냄새 제거법에 관한 문서

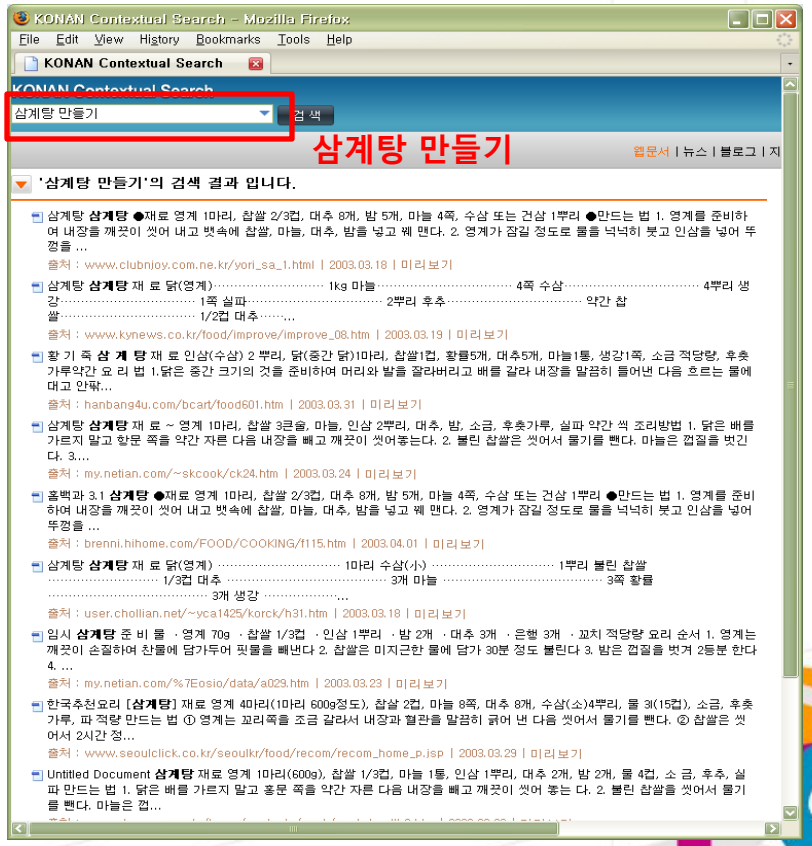


문맥 검색™: 검색 예(3)

구체적 질의 의도에 대해 의미적으로 적합한 단락이 서술된 문서 검색



질의 표현은 다르지만 의미적으로 같은 경우, 동일한 검색 결과 제시



STS2008
Search Technology Summit

www.stskorea.org

Clustered Search

STS
Search Technology Summit

Clustered Search

키워드 수준

구/문장 수준

언어자원과 자연어처리 기술을 활용하여 질의/웹 문서에 대해 의미 분석을 수행하고 질의 의도에 적합한 문서나 정보 제공

검색 결과를 구축된 분류체계를 통하거나 자동적으로 범주화

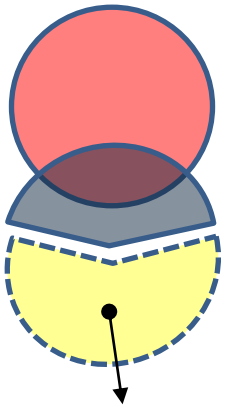
키워드 context 제시

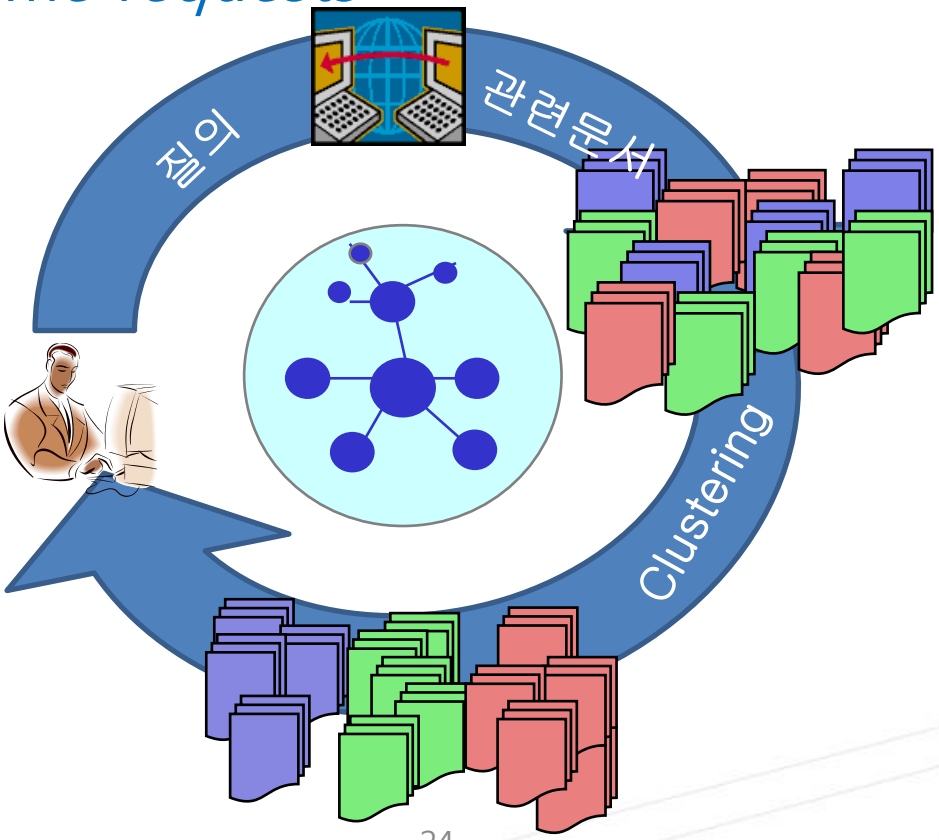
언어자원 | 자연어처리 | 인공지능 | 기계학습

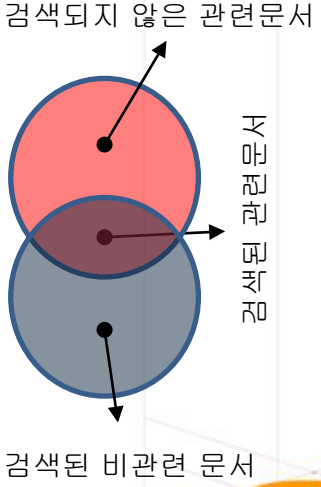
- 이용자의 정보획득 과정의 편의성을 고려
- 일차적인 검색 결과에 대해 범주화된 자료 또는 레이블을 제시함으로써, 원하는 정보에 대한 navigation 과정에 도움.
- 주로 1-2 단어의 키워드 질의에 유용

Clustering in IR

- Cluster Hypothesis (van Rijsbergen, 1971)
  - *"closely associated documents tend to be relevant to the same requests"*



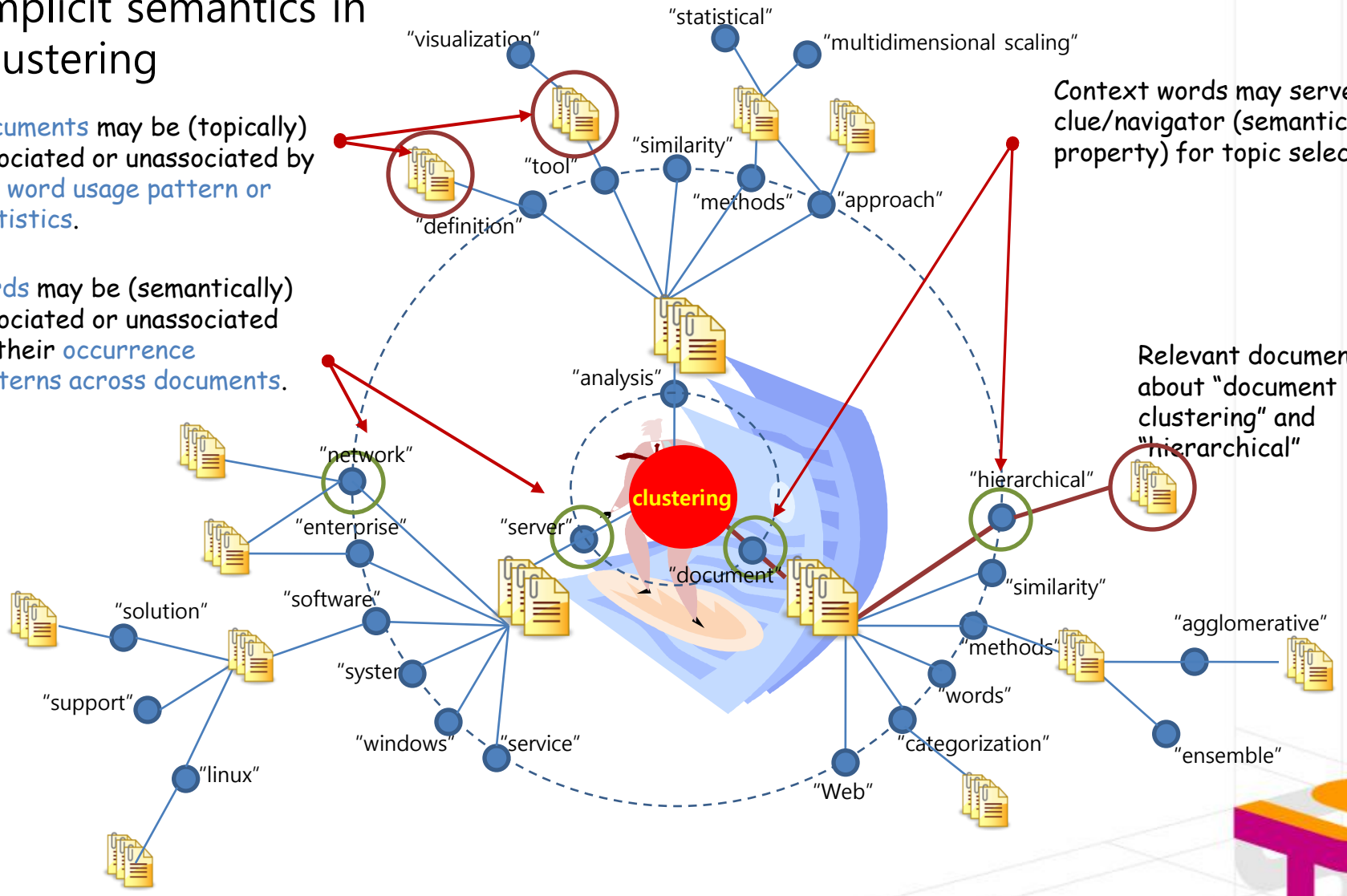




Implicit semantics in clustering

Documents may be (topically) associated or unassociated by the word usage pattern or statistics.

words may be (semantically) associated or unassociated by their occurrence patterns across documents.



Context words may serve as a clue/navigator (semantics or property) for topic selection.

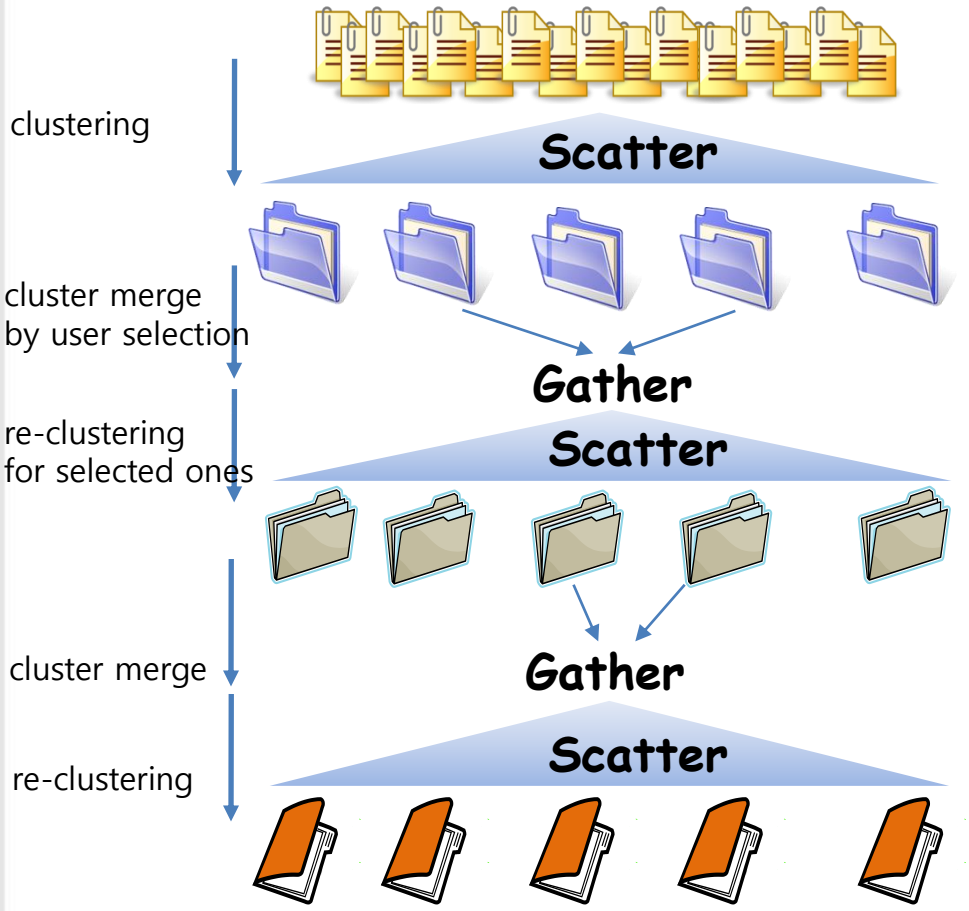
Relevant documents about "document clustering" and "hierarchical"

Context Range in Document Clustering

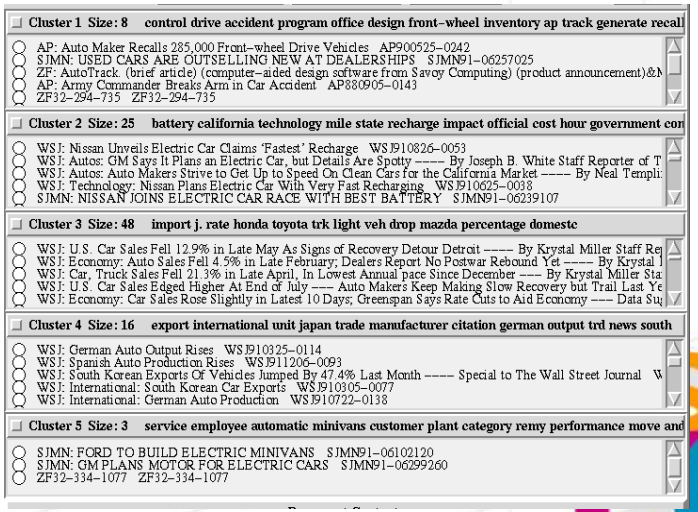
- Clustering with global context
  - Generic, rather query-independent.
  - 질의어를 포함하는 문서의 전체 내용(*global*)에 대한 클러스터링.
  - 문서 내용 전체에 대한 view나 일련의 주제어를 제시.
  - 적용 예: Scatter/Gather.

- Clustering with local context
  - Query-driven.
  - 문서 내에서 질의어의 인접 위치(*local*)의 문맥(paragraph, sentence, phrase)을 대상으로 군집화.
  - 문서 내에서 특정 질의어가 사용된 context 상에서의 view나 속성을 제시.
  - 적용 예: clustering of Web document snippets

Clustering with global context 사례: Scatter/Gather



- 문서의 전체 내용을 기준으로 dynamic clustering
  - 사용자와의 지속적인 상호작용을 통한 반복적 클러스터링
  - 각 문서를 단어들의 집합으로 가정.
  - Fractionation 알고리즘에 의한 클러스터링.
- 각 클러스터는 일련의 주제어와 문서의 제목으로 제시
  - 클러스터 내 단어 빈도 등.



from [Hearst and Pedersen 1996]

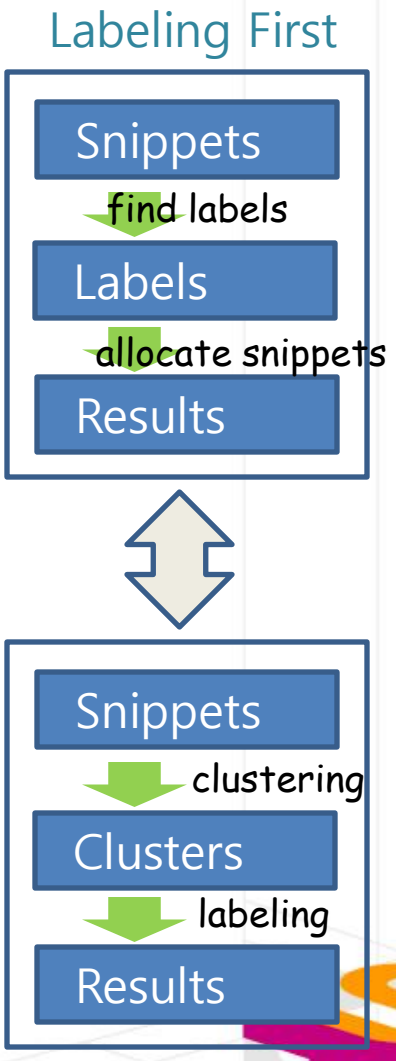
Clustering with local context 사례: STC and LINGO (웹검색 결과 클러스터링)

- 웹문서 snippet에 대한 클러스터링
  - 질의 키워드가 존재하는 문서의 일부에 대해 동적 클러스터링 수행.

Web-snippets

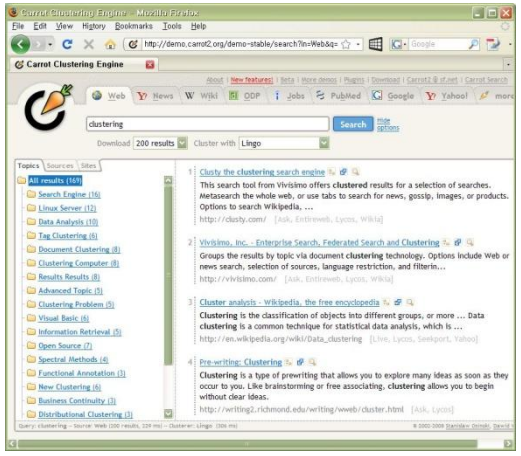
Welcome to Web Services Project @ **Apache**
Kandula - implements WS-Coordination, WS-AtomicTransaction and WS-BusinessActivity protocols based on **Apache** Axis and Axis2. ...

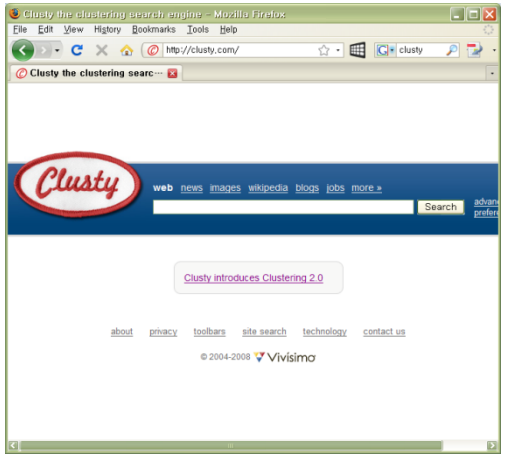
STC	LINGO
빈도수 높은 phrase 추출 & 1차 클러스터링 : **suffix tree**	빈도수 높은 phrase 추출
겹치는 정도가 큰 기본 클러스터들을 병합	클러스터 레이블 추정 : **SVD**, phrase 매칭
최종 클러스터 형성	클러스터에 속하는 문서 동정
	최종 클러스터 형성

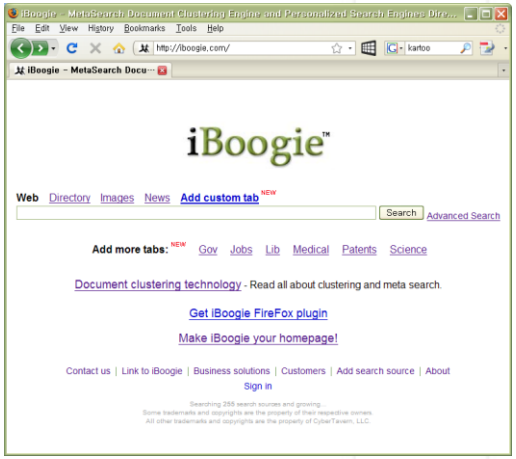


Clustered Search 엔진 사례

Tree View







Map View

감사합니다.