

Semantic Search:

자연어처리와 클러스터링기술의 활용

(주)코난테크놀로지

장 정 호

2008. 9. 2



목 차

- Semantic Search 개요
- 자연언어처리 기술의 활용
- Clustered Search

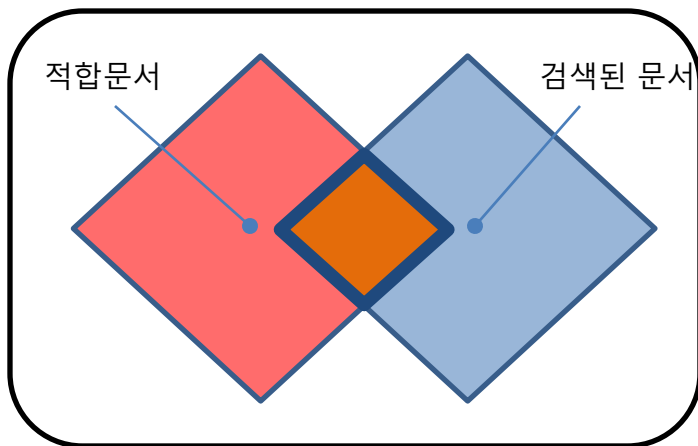
Semantic Search

"Methods of searching web documents *beyond the syntactic level of matching keywords...*

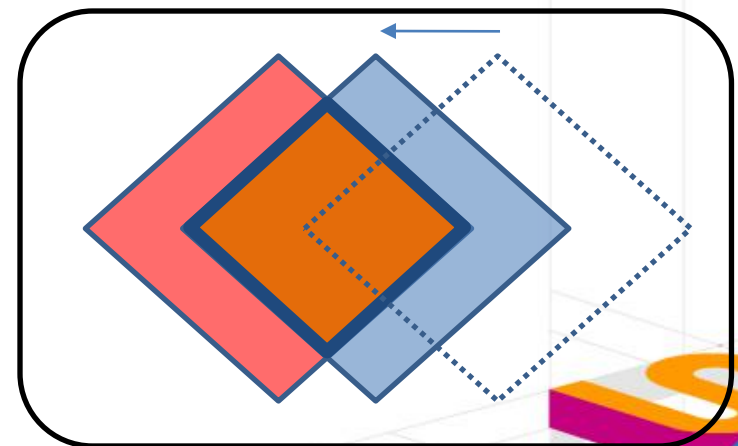
Semantic search - defined as an IR with the capabilities to understand the user's intent and the web's content *at a much deeper, conceptual level.*"

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

키워드 매칭 기반 검색



질의 의미를 파악하고
의미 수준에서 적합 문서 검색

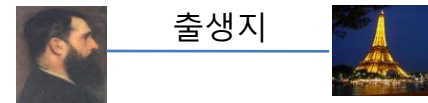
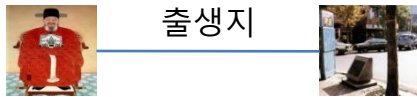



검색결과의 정확도/적합성 향상




Semantics in IR: An Example

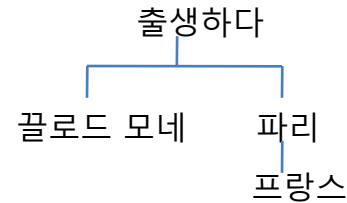
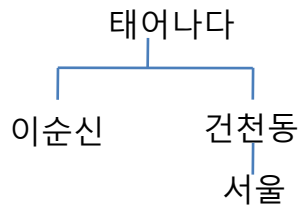
이순신의 출생지는?

모네가 태어난 곳은?



이순신(인물, ) 서울 건천동 (지명, )

클로드 모네(인물, ) 프랑스 파리(지명, ) ~~~~



이순신, 서울, 건천동, 태어나다

클로드, 모네, 프랑스, 파리, 출생하다


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

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

“이순신은 서울의 건천동에서 태어났다”

“클로드 모네는 프랑스의 파리에서 출생했다.”

검색 단계에 따른 의미 정보 응용 예 (semantic + IR)

STAGE	FUNCTIONALITY
<p>Query Construction</p>	<ul style="list-style-type: none"> • 입력 키워드의 의미모호성 해소 파리 → do you mean...? • 키워드 추천 파리(Paris) → 파리교통정보, 파리관광지, 파리여행상품, ... 
<p>Search Process</p>	<ul style="list-style-type: none"> • 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.
<p>Presentation of Search Result</p>	<ul style="list-style-type: none"> • 검색 결과의 조직화 주제/속성별로 분류화된 결과 제시(트리, 맵) • 사용자 피드백 질의문 정제/확장 관련 주제/키워드/리소스 추천

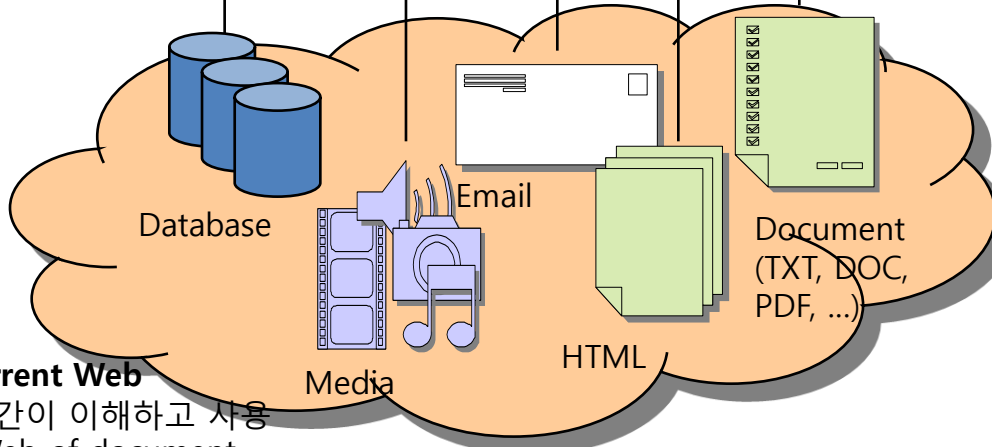
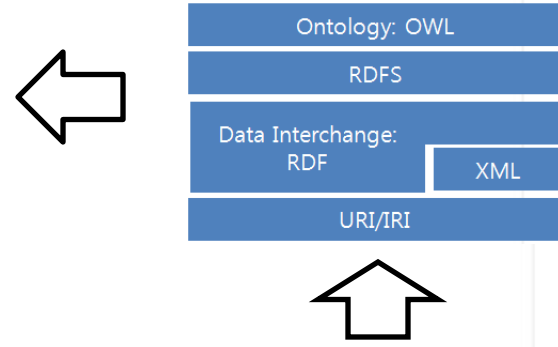
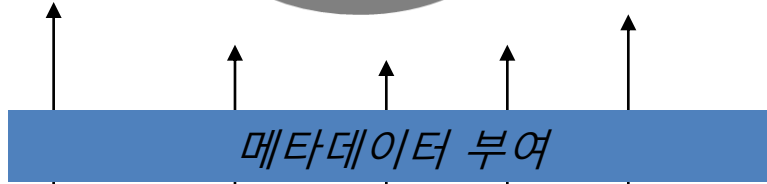
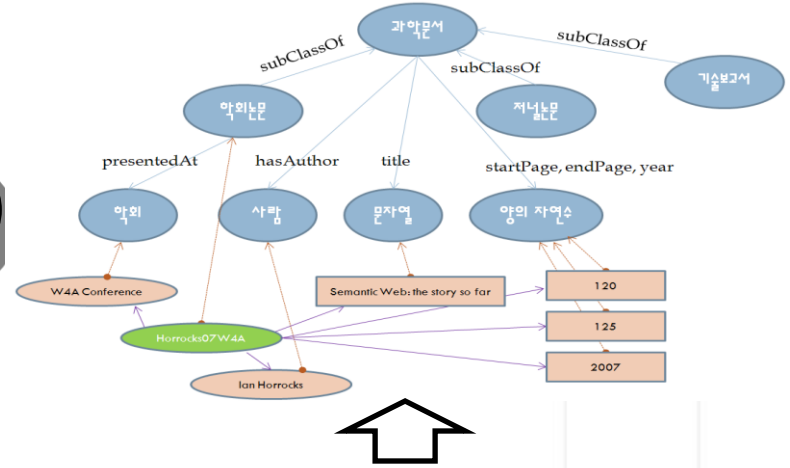
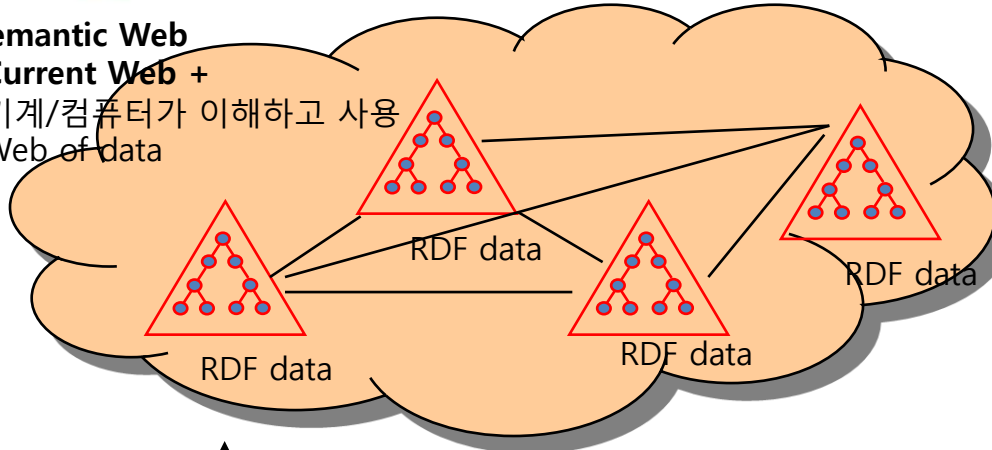
Semantic Sources for Semantic Search

- **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 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.



Search Results Persons: 10 vCards: 1062 Events: 947 Unfinished: 34 | 1 - 10 of about 234039 for ivan herman - 26 sec. (About this page)

- Ivan's blog**
<http://www.ivan-herman.net/> - 47k [Update metadata](#)
- Ivan Herman**


name: Ivan Herman
personal mailbox: <mailto:ivan@w3.org>
homepage: <http://www.ivan-herman.net>

Ivan Herman gives a talk on behalf of the China Office entitled "What is the ... Ben Adida, Elias Torres, and Ivan Herman give a tutorial entitled "RDFa: ...
<http://www.w3.org/People/ivan/> - 23k [Update metadata](#)

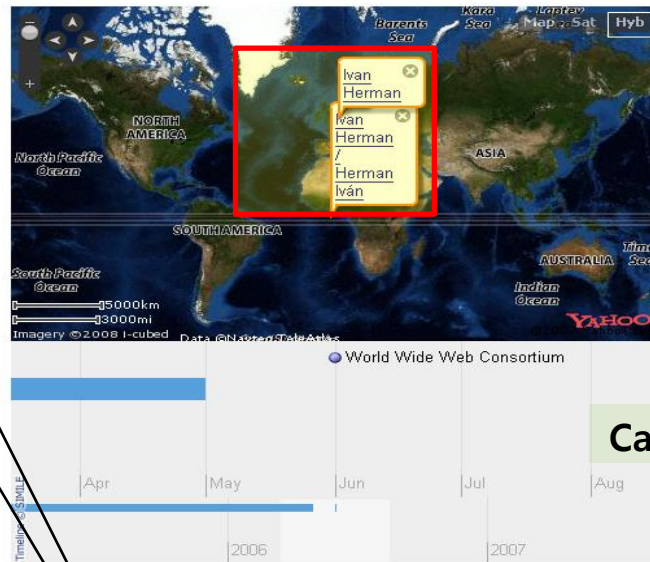
homepage
Ivan's blog
<http://www.ivan-herman.net/> - 47k [Update metadata](#)
- Ivan Herman's Photos**


name: Ivan Herman, Herman Iván, Ivan Herman, Herman Iván
homepage: <http://www.ivan-herman.net>, <http://www.ivan-herman.net>

My Photos. I like traveling and making photos. ... **Iván Herman**, 2008-02-28, ivan@ivan-herman.net The images are licensed under a ...
<http://www.ivan-herman.net/Photos/> - 4k [Update metadata](#)

weblog
Ivan's blog
<http://www.ivan-herman.net/> - 47k [Update metadata](#)

holds account
Ivan Herman | Facebook
 This is **Ivan Herman's** public search listing on Facebook. ...
 Up. — Sign up for Facebook to connect with **Ivan Herman**.
 Already a Member? Login. Name: **Ivan Herman** ...
http://www.facebook.com/p/Ivan_Herman/555188827 - 12k
[Update metadata](#)



Map: Ivan Herman, Ivan Herman, Herman Iván

Calendar: World Wide Web Consortium, Apr, May, Jun, Jul, Aug, 2006, 2007

Map 데이터와 연동

Calendar 정보

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

참조: Yahoo! microSearch



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 Yīlóngyī*; Wade-Giles: *Tai-pei I-lung-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 m ² (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

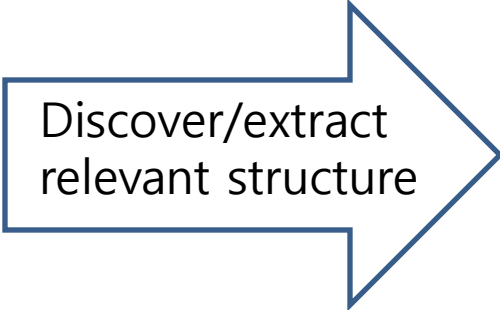
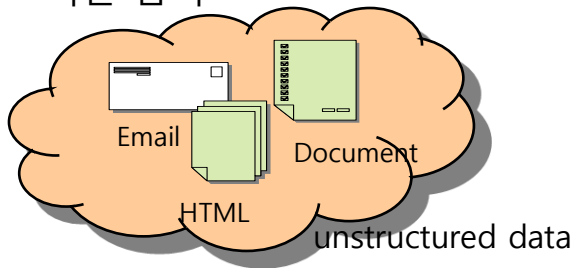
101

constructed_year

1999-2004

Exploiting Implicit Semantics

질의 키워드의 단순 매칭에 의한 검색



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



NLP 기반 semantic search



자연어 분석 과정



NLP에 의한 문장 분석 예

모네는 프랑스의 파리에서 태어났다.

단어(어절) 추출
문장 부호 분리
숫자/특수문자열 처리

전처리

w1: 모네는 w2: 프랑스의 w3: 파리에서 w4: 태어났다

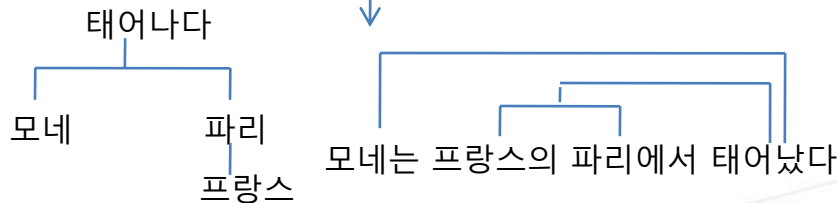
결합제약규칙
원형복원규칙
시스템사전

형태소 분석

모네/NC+는/PP 프랑스/NR+의/PN 파리/NC+에서/PP
태어나/VV+았/EP+다/EF

문법규칙
분석알고리즘
어휘사전

구문 분석



모네의 출생지 언급

파리: Paris 태어나다(모네, 파리)

의미 분석

시소러스
의미사전
통계정보



언어자원과 NLP기반 분석에 기초한 semantic search

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

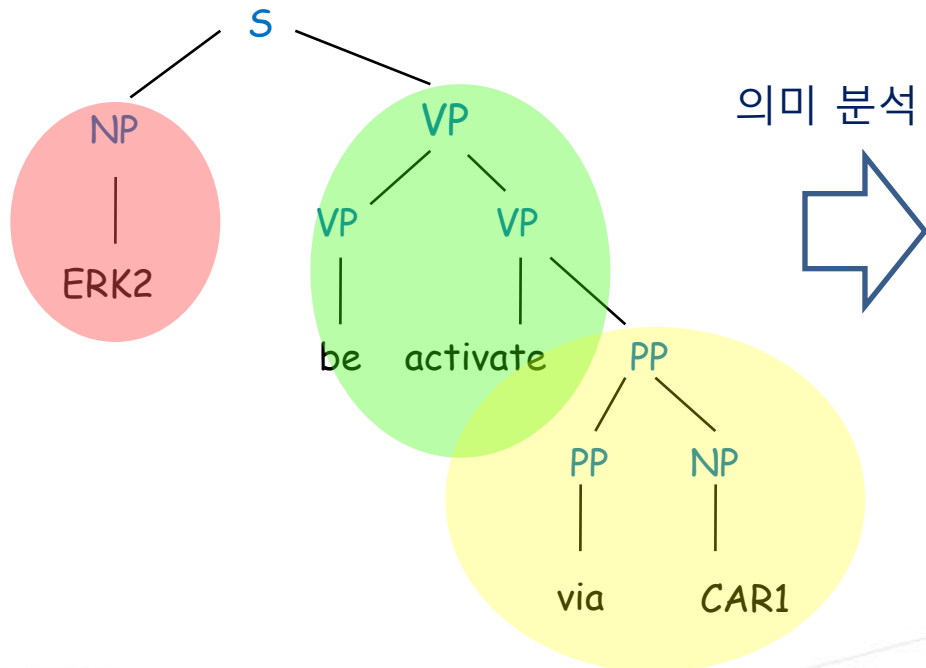


형태소 분석

... ERK2/**NN** is/**VBZ** activated/**VCN** via/**IN** CAR1/**NN**...



구문 분석



의미 분석



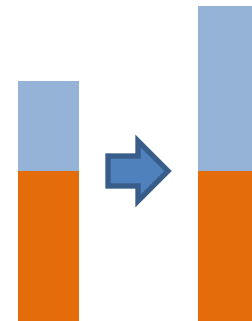
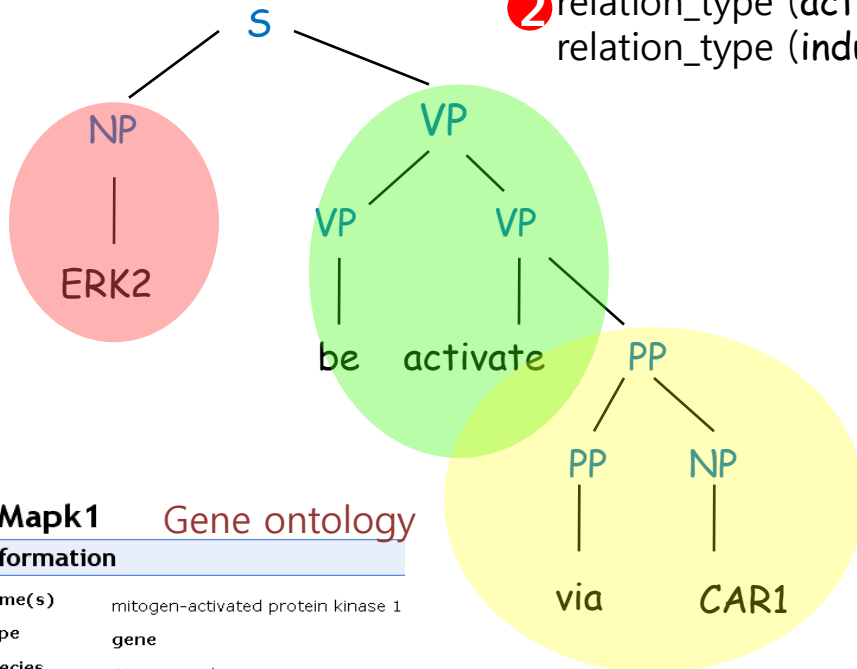
ERK2: gene
CAR1: gene
rel_activation(CAR1, ERK2)



Query: What induces MAPK1?

Word ontology

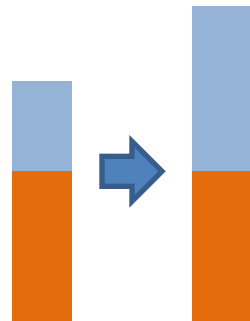
② relation_type (activate) = 'activation'
relation_type (induce) = 'activation'



recall 향상

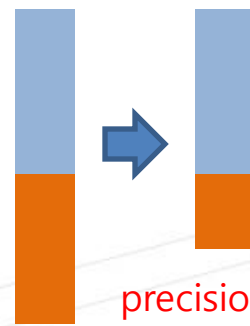
① Mapk1 Gene ontology Information

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

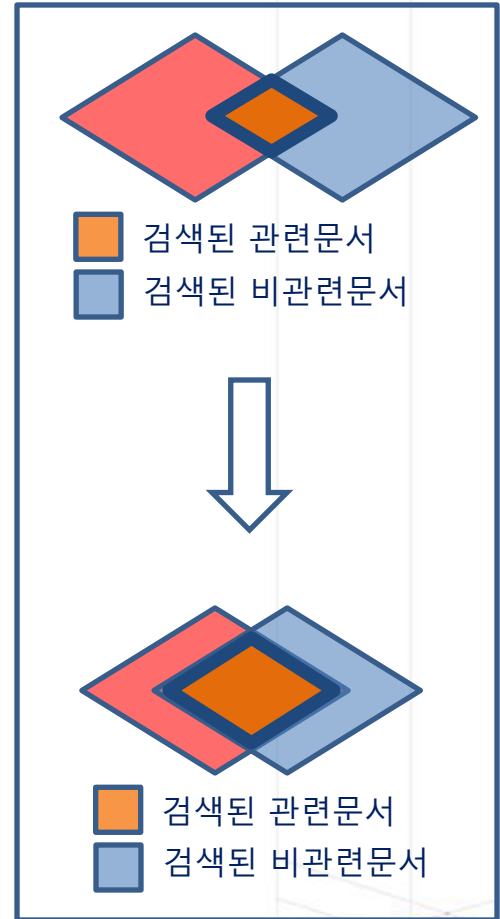


recall 향상

③ rel_activation(CAR1, ERK2)



precision 향상



검색된 관련문서
검색된 비관련문서

검색된 관련문서
검색된 비관련문서

NLP 기반 Semantic Search 엔진 사례

❑ Cognition

"... technology that understands word and phrase meanings within context in modern computer applications..." (from <http://www.cognition.com>)



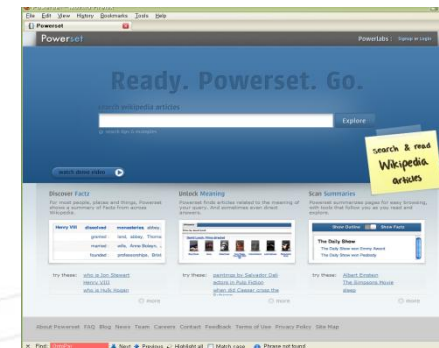
❑ Hakia

"... a semantic search engine that brings relevant results based on concept match rather than keyword match or popularity ranking." (from <http://www.hakia.com>)



❑ PowerSet

"... applying its natural language processing to search, aiming to improve the way we find information by unlocking the meaning encoded in ordinary 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



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

KONAN Contextual Search - Mozilla Firefox

File Edit View History Bookmarks Tools Help

KONAN Contextual Search

시드니 검색

시드니 업문서 | 뉴스 | 블로그 |

적중검색

- 시드니 역사**
시드니는 호주에서 가장 오랜 역사를 가지고 있으며, 뉴사우스웨일스 주의 주도이다. 호주 최대의 도시로 1770년 제임스 쿡 선장이 이끄는 탐험대에 의해 시드니 할란이 최초로 발견되었다. 시드니는 세계 3대 미항인 시드니항과 코발트 빛 바다와...
출처 : www.ojin.co.kr/oversear/sydney/sydney01.htm | 2003.03.28 | 미리보기
- 시드니 소개**
...마운틴은 탐험가들에 의하여 알려 지게 되었고 시드니는 정부로부터 NEW SOUTH WALES 의 서부 도시로서 명명하게 되었다. 1850년 빅토리아와 서부 시드니에서 금이 발견 되어 있을 때 새로운 정착자들은 금을 발견하기...
출처 : tourdate.co.kr/c/c_1_sydney.htm | 2003.03.21 | 미리보기
- 시드니 위치**
시드니는 세계 3대 미항인 시드니항과 코발트 빛 바다와 어우러지는 오페라하우스 볼 수 있는 아름다운 도시로 호주의 경제·문화의 중심지로 남위 34°에 위치하며 온대성 기후대에 속하나 해양성기후의 영향으로 여름...
출처 : hanabank.interparktravel.com/tour_information/southpacific/oceania/Australia_infosub_07.asp | 2003.03.29 | 미리보기
- 시드니는 어떤가요?**
세계에서 가장 아름다운 항구에 세워진 시드니는 2000년 시드니 올림픽을 개최하면서 보다 더 국제적인 도시로 발돋움 하고 있으며 전세계 도시명가에서도 급간에 연속 3년간 최고의 도시로 평가받고 있다. 호주의 경제...
출처 : touryours.co.kr/html/region/ose_sydney_1.jsp | 2003.03.17 | 미리보기
- 시드니는 무엇으로 유명한가요?**
시드니는 세계 미항 중의 하나로 알려져지는 아름다운 항구를 끼고 있는 해변, 문화 행사와 오페라 하우스, 허버리리지의 건축물로 유명합니다. 시드니는 뉴 사우스 웨일스주의 많은 명소로 연결되는 관문이기도 합니다. 호주 원주민들의 바...
출처 : icbc.net/main01_aus03.html | 2003.03.26 | 미리보기

▼ '시드니'의 검색 결과입니다.

- 호주 관광지 안내입니다. ...마운틴은 탐험가들에 의하여 알려 지게 되었고 시드니는 정부로부터 NEW SOUTH WALES 의 서부 도시로서 명명하게 되었다. 1850년 빅토리아와 서부 시드니에서 금이 발견 되어 있을 때 새로운 정착자들은 금을 발견하기 위하여 시드니로 몰렸듯이 ...
출처 : tourdate.co.kr/c/c_1_sydney.htm | 2003.03.21 | 미리보기
- 국토청 유한닷컴 > 특집 New South Wales주, 시드니 시드니는 호주 최대의 도시로서 전체 인구의 25%에 가까운 약 400만 명이 살고 있는 시드니지역을 중심으로 한 NSW는 호주의 중심이라고 말할 수 있다. 물가는 다른 지역 에 비해 비싸지만 많은 학교 등이 집중되어 있다...
출처 : www.goodmorningsuhak.com/info_021.htm | 2003.03.21 | 미리보기
- Untitled Document > O2(아이시나 한글) 해외하는 도시 시드니 특히 사할 인연에서 시드니까지 직할로 운항합니다. 학생 편도, 학생 반도, 학생 2인 1실, 학생 3인 1실은 국제학생들, 국제학생증 소지자 오릅니다. 적용 기간 11/1~11/20 3/1~3/31 110 60 11/21~12/14 2/1~2/28 120 70 1...
출처 : kises.com.kr/air_aus.php3 | 2003.03.20 | 미리보기
- 배낭여행/해외/무엇/여행/준비/여행/준비/여행/준비 시드니 시드니에는 사해에서 이용할 수 있는 학교투어스가 2가지 종류

KONAN Contextual Search - Mozilla Firefox

File Edit View History Bookmarks Tools Help

KONAN Contextual Search

지방간 검색

지방간 업문서 | 뉴스 | 블로그 |

적중검색

- 지방간은 어떤 질병인가?**
내과 간담도계 -지방간 지방간이란 지방간은 간에 염증이 간세포의 파괴없이 간세포에 기름이 끼어 간 전체가 커지는 질환이다. 원 인 건강진단을 받은 우리나라 성인남성의 약30%, 성인여...
출처 : www.hanbangmoim.com/disease/14/14b_a02.htm | 2003.03.21 | 미리보기
- 지방간의 증상**
지방간 이 증상 이 지방간은 특별한 증상을 느끼지 못하는 경우가 많으며 증상이 있어도 지방간을 의심할 수 있는 특이한 것은 없고 간질환시의 일반적인 증상인 증상인 피로감이나 식욕감소 등이 나...
출처 : www.kuri365.com/life/life05_01_open8.html | 2003.03.27 | 미리보기
- 지방간의 원인**
지방간은 주로 과다한 음주와 비만으로 발생한다. 고지 혈증 당뇨병 갈색성기능항진증 등의 내분비질환, 부신피질호르몬 또는 여성호르몬 등 의 과다사용도 원인이 될 수 있다. 한의학에서는 간중증 주상증의 범주에서 지방간을 해석하고 있다. ...
출처 : www.hanbangmoim.com/disease/14/14b_a02.htm | 2003.03.21 | 미리보기
- 지방간의 예방법**
지방간을 예방하기 위해서는 과음과 과식을 피하고, 지방간 발생의 원인이 될 수 있는 당뇨병과 고지혈증을 조기에 발견하여 치료 합니다. 규칙적인 운동을 통해 정상체중으로 유지하고, 기름기가 많은 음식과 감미식...
출처 : www.rodemclinic.co.kr/digest/sub4_2.asp | 2003.03.31 | 미리보기

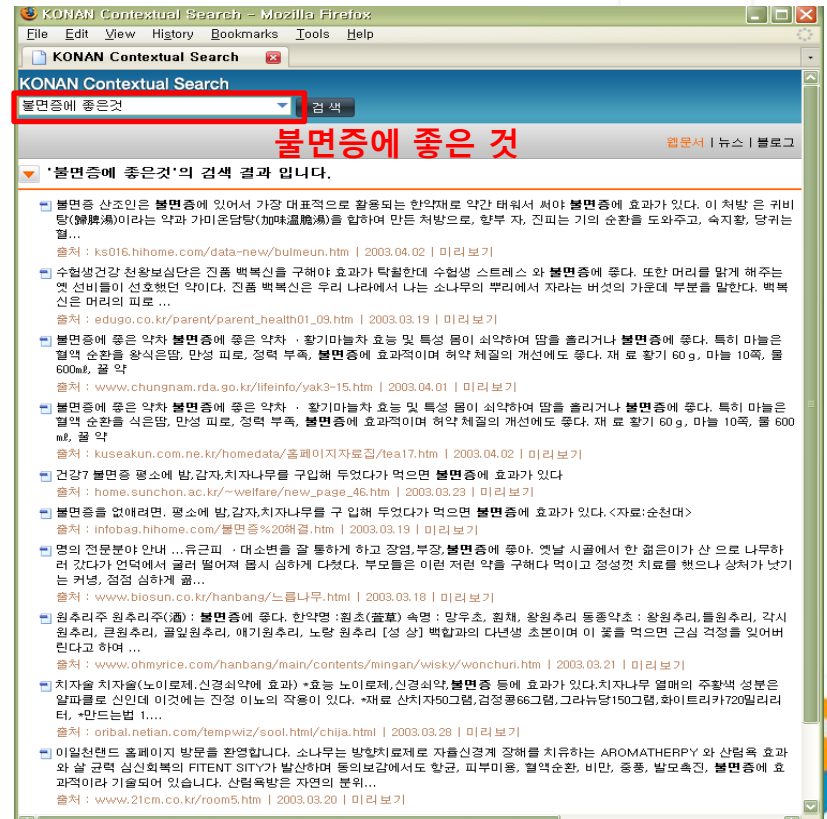
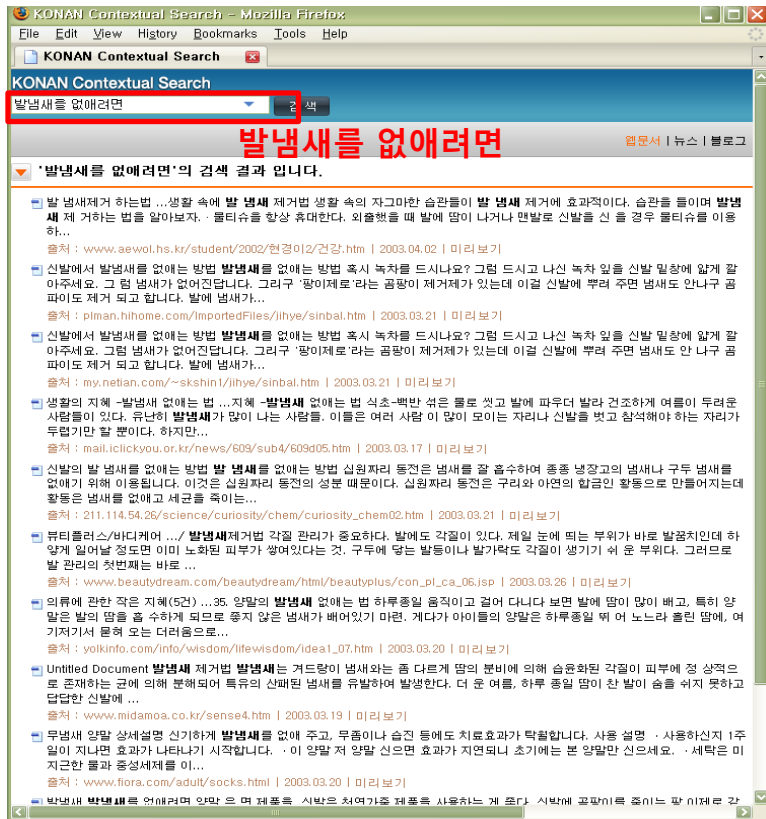
▼ '지방간'의 검색 결과입니다.

- Untitled Document ...간담도계 -지방간 지방간이란 지방간은 간지지방간은 주로 과다한 음주와 비만으로 발생한다. 고지 혈증 ...
출처 : www.hanbangmoim.com/disease/14/14b_a02.htm | 2003.03.21 | 미리보기
- 지방간 (Fatty liver) 지방간 (fatty liver) 지방간이란 간의 지방대사장애로 증성지방과 그의 전구체인 지방산이 간세포에 축적된 상태를 말한다...당뇨병 등 과 연관되어 유발되는데, 이에 따라 지방간의 종류에는 알코올성 지방간, 비알코올성 지방간, 비만성 지방간
출처 : nopain365.com/fattyliver_01.html | 2003.03.19 | 미리보기
- 김치만 내과의원 :::: 지방간 (Fatty Liver) 지방간이란 증성 지방(triglyceride)이 간세포 속에 축적된 상태를 말한다. 약간의 지방간 소견은 지방간은 그 원인을 제거하면 대부분이 정상체로 회복되지만, 알코올성 지방간인 경우 음주를 계속하면 간경변 등, 간암 등
출처 : www.kimsmed.com/sub3.html | 2003.03.29 | 미리보기
- 안녕하세요 ~~~ 최 차해 병원입니다...(Choi's Hospital) 지방간이란 트리글리세라이드(TG)라는 지방 이 간세포 속에 축적되는 질환으로 간 무게 의 5% 이상을 이 트리글리세라이드가 차지지방간의 원인은 비만, 과음, 당뇨병, 고지혈증, 스테로이드 등의 약제로 인한 지방간 등이 있다. 보통 지방간의 간
출처 : yeannm.com.kr/negwa3_12.htm | 2003.03.18 | 미리보기
- 지방간 지방간 간경변증 지방간(Fatty liver) 지방간은 간에 지질이 이상적으로 많이 축적되어 있는 질병을 말합니다. 이 ... 만 이 상의 지방공포가 보이는 상태가 올 때를 지방간이라고 합니다. 원인은 영양장애(기아, 단백질 부족, 비타민 결핍, 콜
출처 : foodlove.pe.kr/foodlee/145_1.htm | 2003.03.19 | 미리보기

문맥 검색™: 검색 예(2)

구체적 질의 의도에 대해 의미적으로 적합한 문장이 서술된 문서 검색

발냄새 제거법에 관한 문서

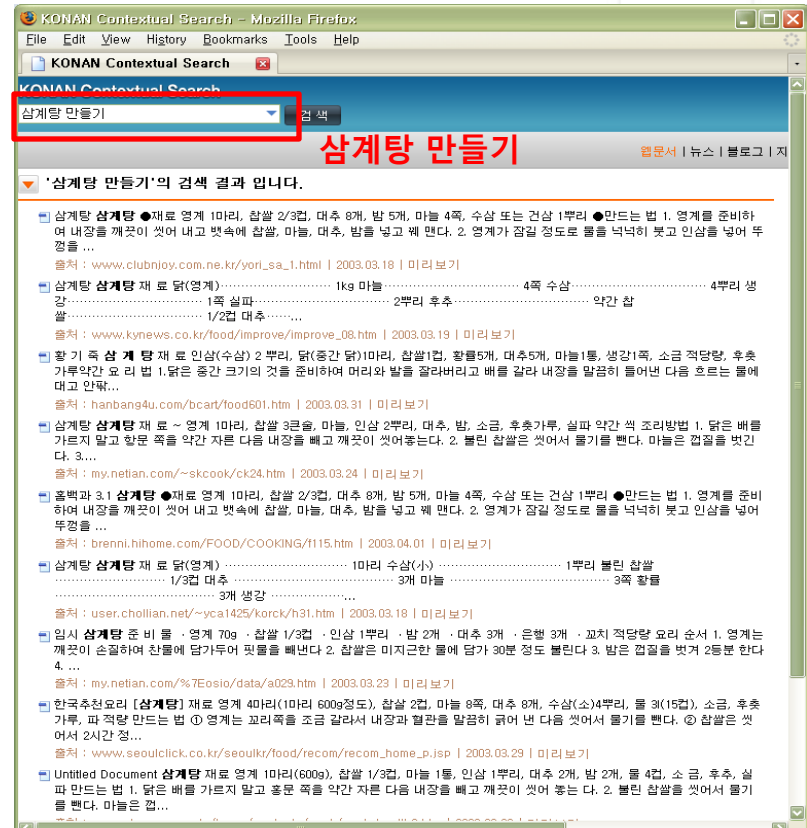
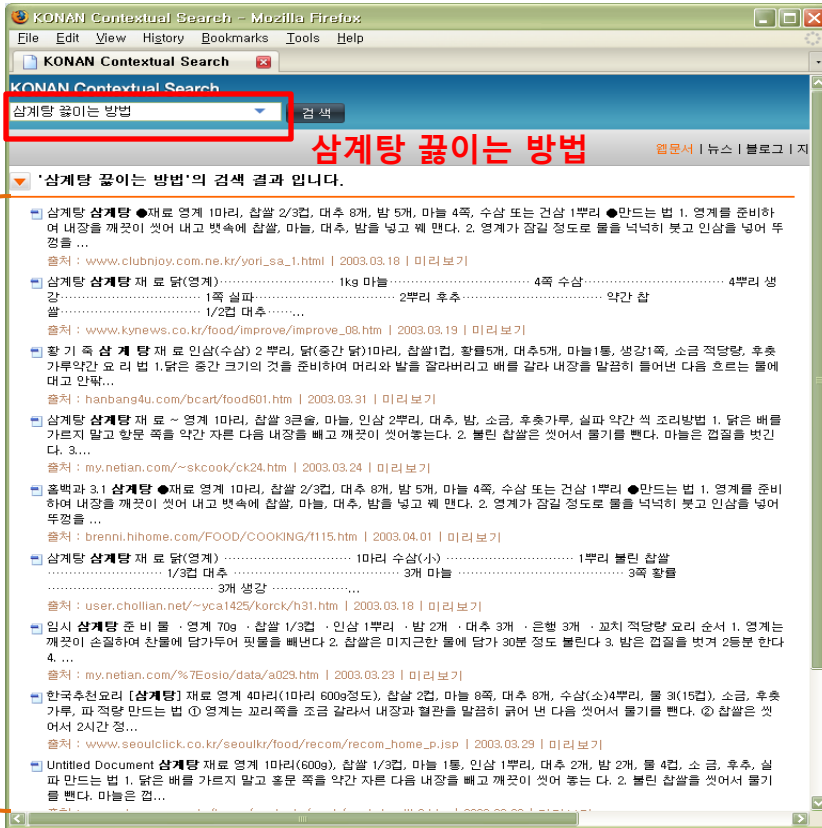


문맥 검색™: 검색 예(3)

구체적 질의 의도에 대해 의미적으로 적합한 단락이 서술된 문서 검색

질의 표현은 다르지만 의미적으로 같은 경우, 동일한 검색 결과 제시

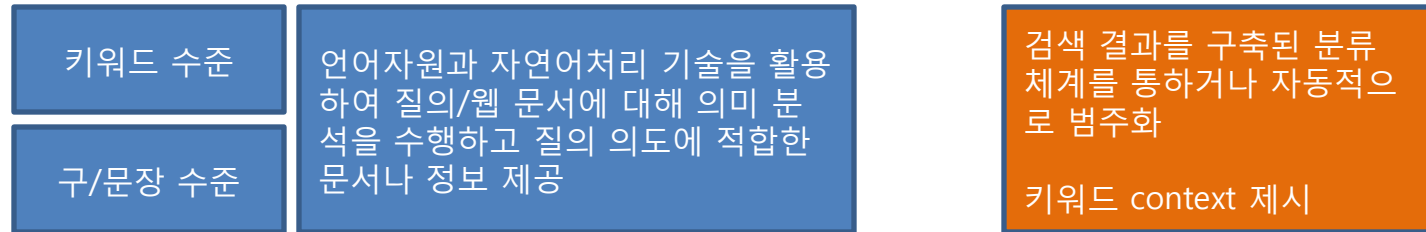
삼계탕 요리법에 관한 문서



Clustered Search



Clustered Search

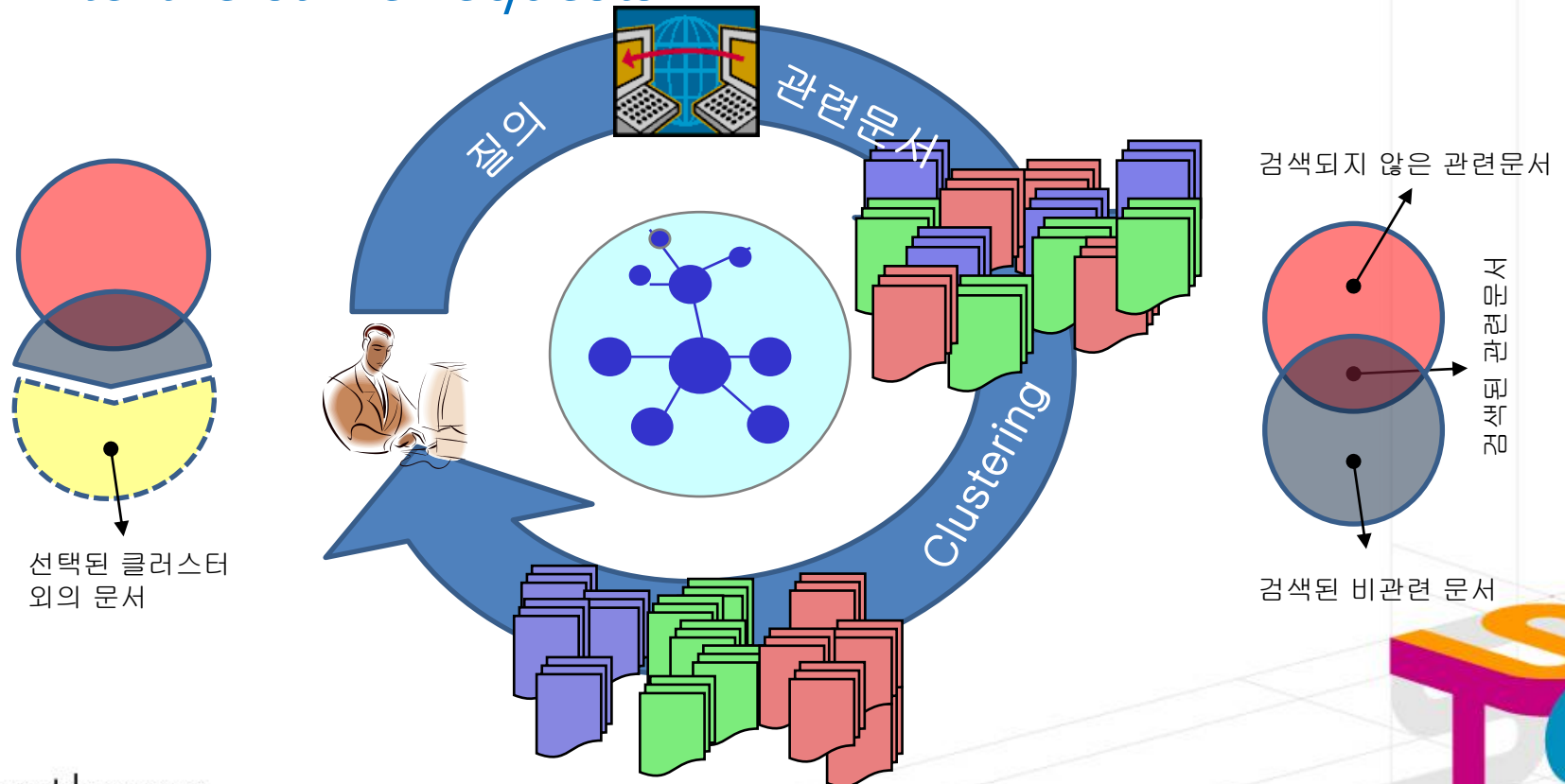


- 이용자의 정보획득 과정의 편의성을 고려
- 일차적인 검색 결과에 대해 범주화된 자료 또는 레이블을 제시함으로써, 원하는 정보에 대한 navigation 과정에 도움.
- 주로 1-2 단어의 키워드 질의에 유용



Clustering in IR

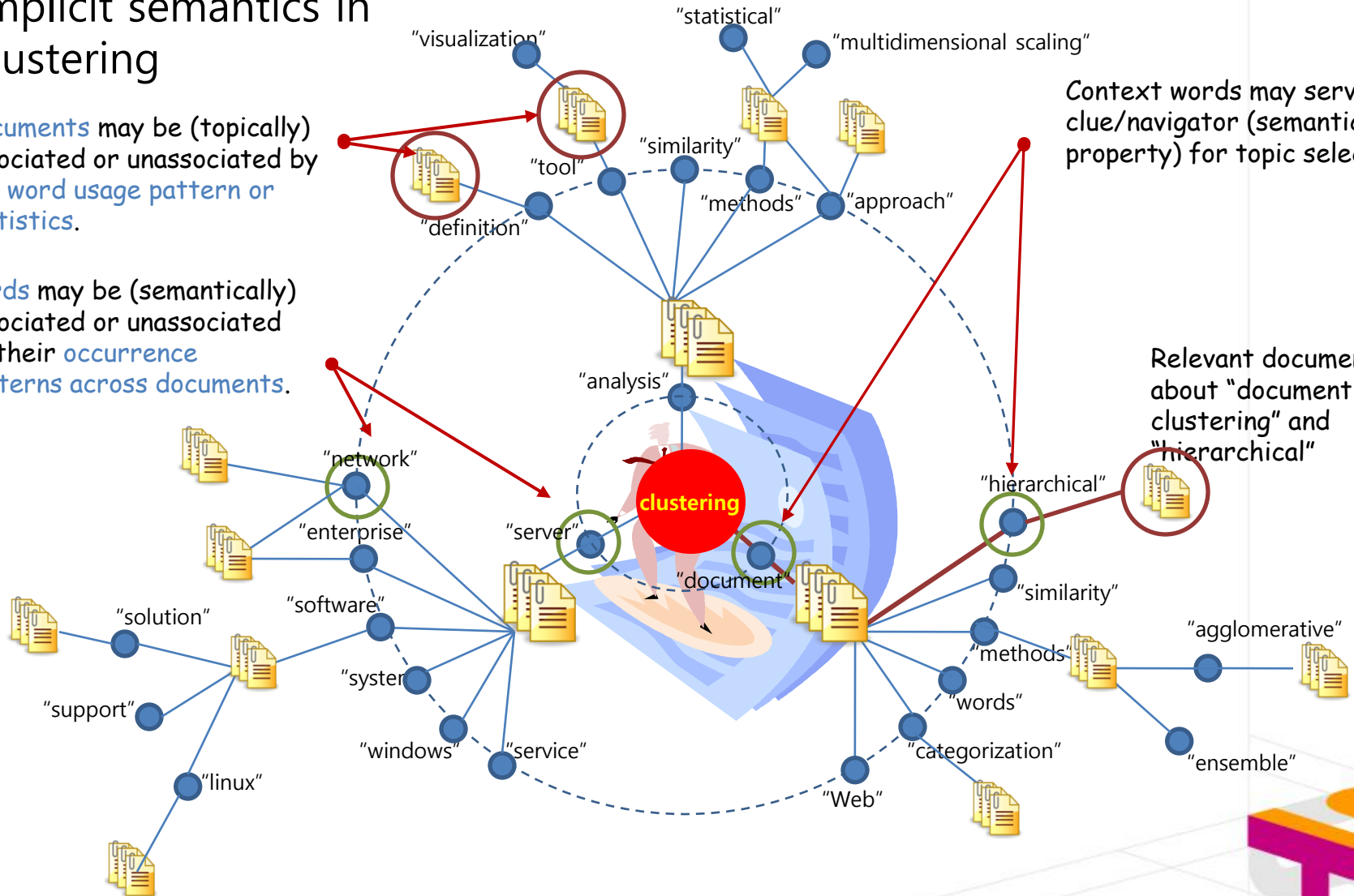
- Cluster Hypothesis (van Rijsbergen, 1971)
 - "*closely associated documents tend to be relevant to the same requests*"



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"

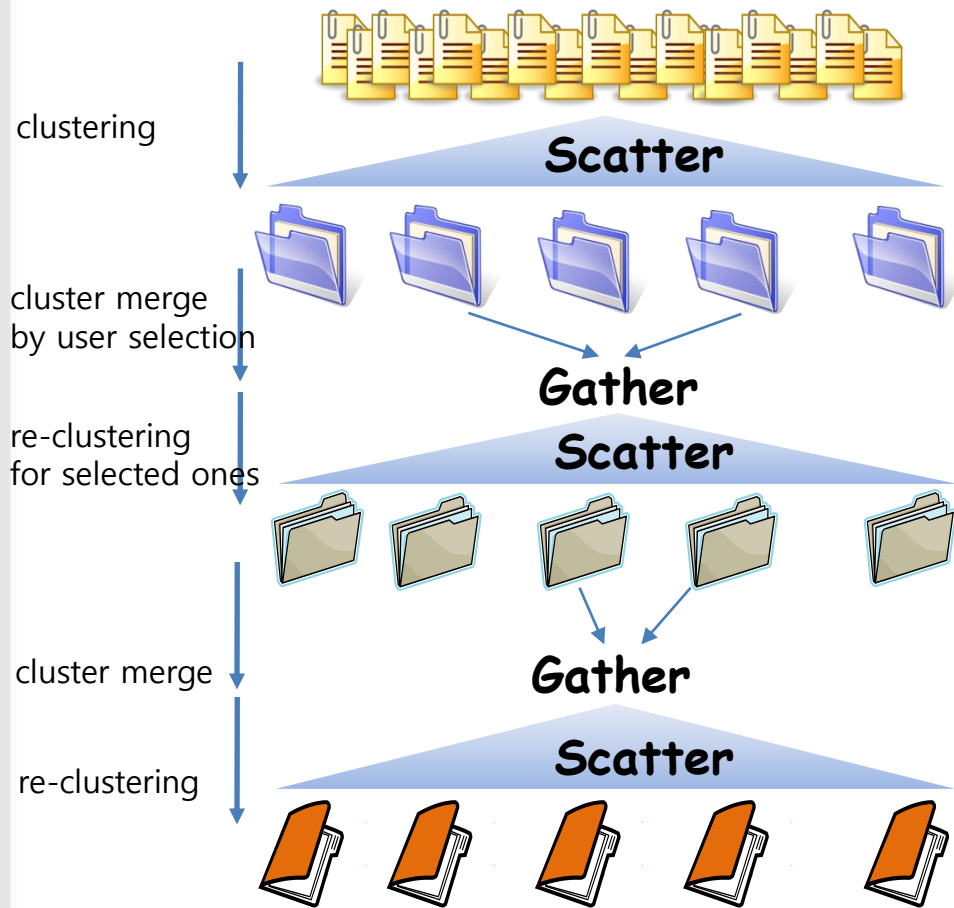


Context Range in Document Clustering

- 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.

Cluster 1 Size: 8	control drive accident program office design front-wheel inventory ap track generate recall
<input type="radio"/>	AP: Auto Maker Recalls 285,000 Front-wheel Drive Vehicles AP900525-0242
<input type="radio"/>	SJMN: USED CARS ARE OUTSELLING NEW AT DEALERSHIPS SJMN91-06257025
<input type="radio"/>	ZF: AutoTrack (brief article) (computer-aided design software from Savoy Computing) (product announcement)&#
<input type="radio"/>	AP: Army Commander Breaks Arm in Car Accident AP880905-0143
<input type="radio"/>	ZF32-294-735 ZF32-294-735
Cluster 2 Size: 25	battery california technology mile state recharge impact official cost hour government con
<input type="radio"/>	WSJ: Nissan Unveils Electric Car Claims 'Fastest' Recharge WSJ10826-0053
<input type="radio"/>	WSJ: Autos: GM Says It Plans an Electric Car, but Details Are Spotty ---- By Joseph B. White Staff Reporter of T
<input type="radio"/>	WSJ: Autos: Auto Makers Strive to Get Up to Speed On Clean Cars for the California Market ---- By Neal Temp
<input type="radio"/>	WSJ: Technology: Nissan Plans Electric Car With Very Fast Recharging WSJ10025-0038
<input type="radio"/>	SJMN: NISSAN JOINS ELECTRIC CAR RACE WITH BEST BATTERY SJMN91-06239107
Cluster 3 Size: 48	import j. rate honda toyota trk light veh drop mazda percentage domestic
<input type="radio"/>	WSJ: U.S. Car Sales Fell 12.9% in Late May As Signs of Recovery Detour Detroit ---- By Krystal Miller Staff Re
<input type="radio"/>	WSJ: Economy: Auto Sales Fell 4.5% in Late February; Dealers Report No Postwar Rebound Yet ---- By Krystal M
<input type="radio"/>	WSJ: Car, Truck Sales Fell 21.3% in Late April, in Lowest Annual Pace Since December ---- By Krystal Miller Sta
<input type="radio"/>	WSJ: U.S. Car Sales Edged Higher At End of July ---- Auto Makers Keep Making Slow Recovery but Trail Last Ye
<input type="radio"/>	WSJ: Economy: Car Sales Rose Slightly in Latest 10 Days; Greenspan Says Rate Cuts to Aid Economy ---- Data Su
Cluster 4 Size: 16	export international unit japan trade manufacturer citation german output trd news south
<input type="radio"/>	WSJ: German Auto Output Rises WSJ10375-0114
<input type="radio"/>	WSJ: Spanish Auto Production Rises WSJ11206-0093
<input type="radio"/>	WSJ: South Korean Exports Of Vehicles Jumped By 47.4% Last Month ---- Special to The Wall Street Journal W
<input type="radio"/>	WSJ: International: South Korean Car Exports WSJ10305-0077
<input type="radio"/>	WSJ: International: German Auto Production WSJ10722-0138
Cluster 5 Size: 3	service employee automatic minivans customer plant category reny performance move and
<input type="radio"/>	SJMN: FORD TO BUILD ELECTRIC MINIVANS SJMN91-06102120
<input type="radio"/>	SJMN: GM PLANS MOTOR FOR ELECTRIC CARS SJMN91-06299260
<input type="radio"/>	ZF32-334-1077 ZF32-334-1077

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

빈도수 높은 phrase 추출
& 1차 클러스터링
: **suffix tree**

겹치는 정도가 큰 기본 클러스터들을 병합

최종 클러스터 형성

LINGO

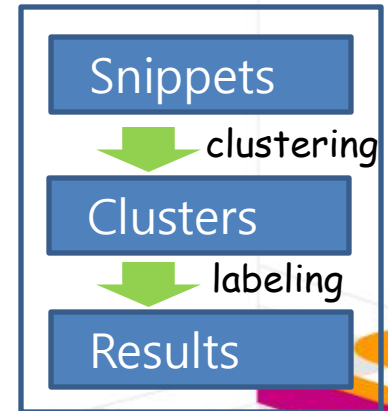
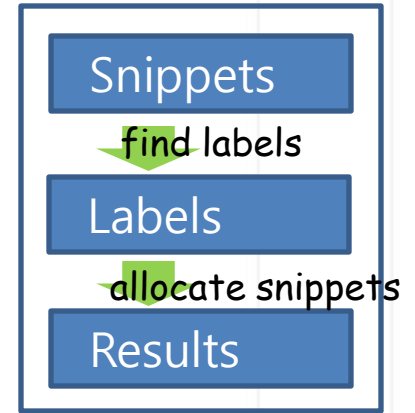
빈도수 높은 phrase 추출

클러스터 레이블 추정
: **SVD**, phrase 매칭

클러스터에 속하는 문서 동정

최종 클러스터 형성

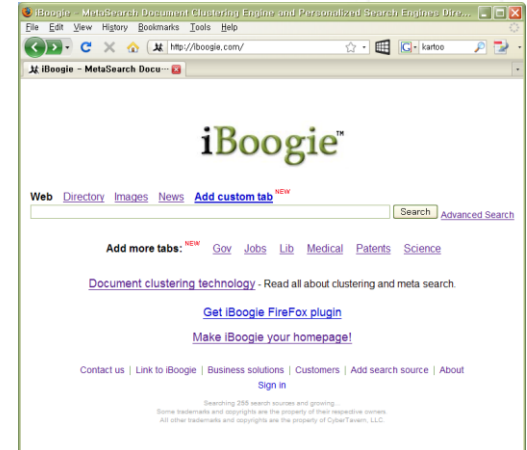
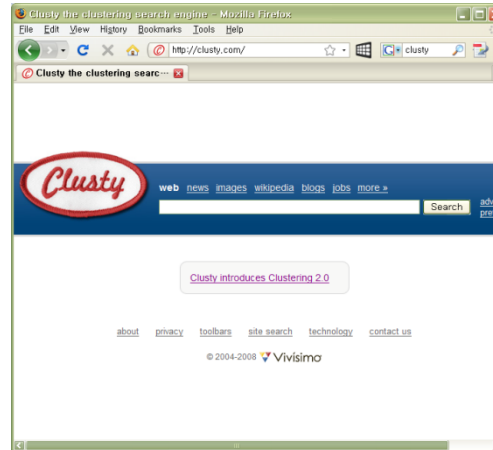
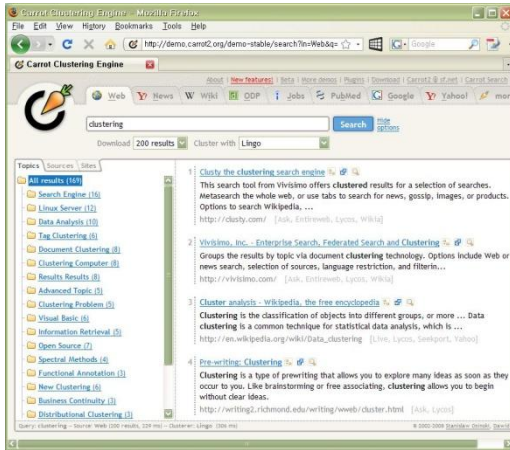
Labeling First



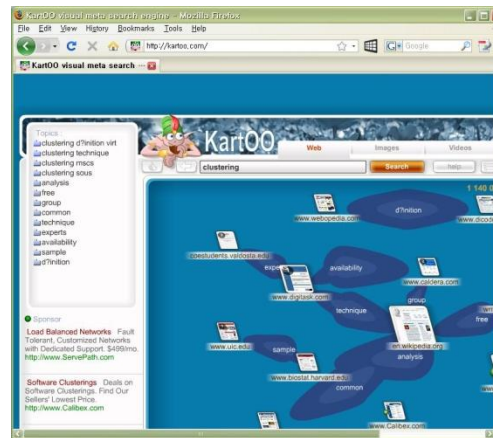
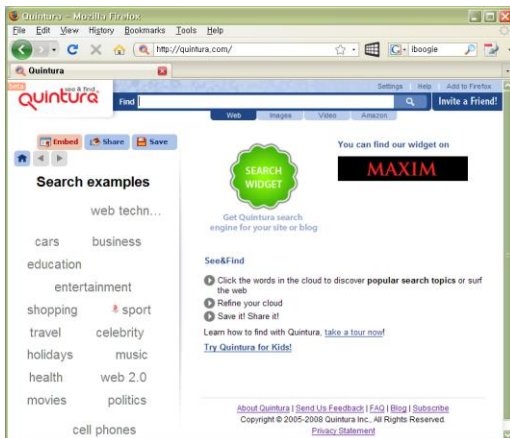
Clustering First

Clustered Search 엔진 사례

Tree View



Map View



감사합니다.

