



# Web Search Engine

국민대학교 소프트웨어학부

강 승 식

# 차 례

---

- 웹 검색엔진
  - 국내외 검색엔진
  - 웹의 특성 및 사용자 특성
- 웹 검색엔진 issues
  - Web spider(crawler)
  - Ranking : 문서 연관성 기법
    - PageRank, HITS

# 검색엔진 개발(국외)

---

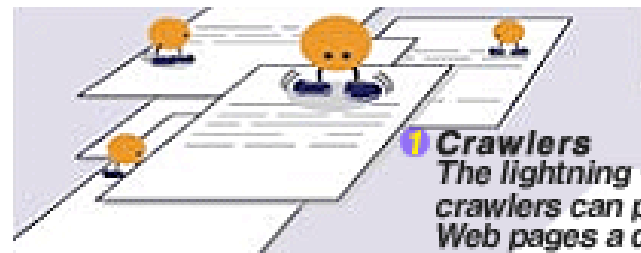
- Lycos : CMU의 연구 프로젝트(1994)
- Excite : Stanford 대학원생
- OpenText : 워싱턴 대학
- HotBot
  - U.C. Berkley의 검색엔진을 발전시킴
- Altavista : DEC(1995)
- Google : Stanford 박사과정 학생
- InkTomi, Northernlight 등
- Ask Jeeves(1997.4)
- Answerbus.com(2001)

# 웹 검색엔진 서비스(국내)

---

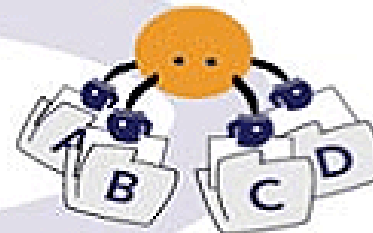
- Simmani(1996.3) : 한글과 컴퓨터
- 까치네(1996) : 대구대 동아리
- 한글 Yahoo(1997.9)
- Naver(1998.1) : 삼성 SDS
- Altavista(1998.5)
- Lycos(1999.7)
- Empas(1999.11)
- HanMir, 와카노, Paran
- Google(2000.9)

# <http://kr.wisenut.com/>




**1 Crawlers**  
The lightning -fast WISEnut crawlers can process many Web pages a day with just a handful of servers.

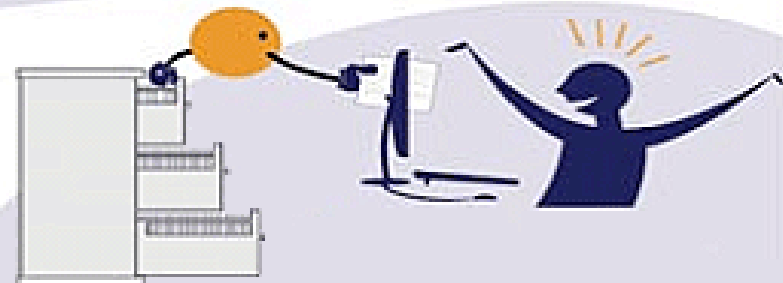
**2 Indexer**  
The WISEnut indexer is just as fast. By the full launch, the WISEnut launch database will contain more web pages than any existing search engine



**3 Ranking System**  
Our unique ranking algorithm will generate relevant results across a wide range of search queries at a faster rate than our competitors



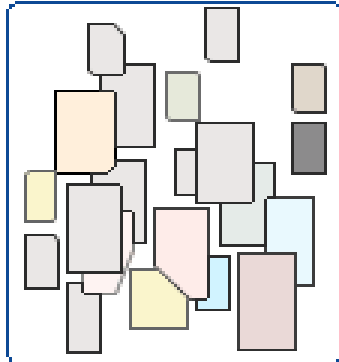
	A	B	C	D
1	▲	—	⤴	👤
2	●	📄	⤴	👤
3	■	📄	⤴	👤
4	●	📄	⤴	👤



**4 Searcher**  
The WISEnut search engine returns more results faster. Clustering by samifarity and domain name provides the widest range of result.

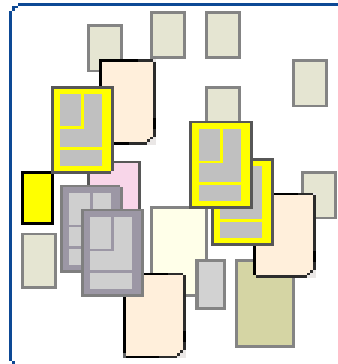
# <http://kr.wisenut.com/>

Web Pages



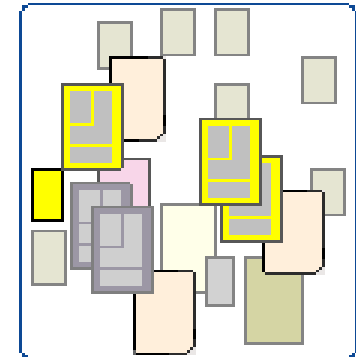
Comprehensive  
Crawling

Large Index DB



Sort  
Refresh  
Cycle

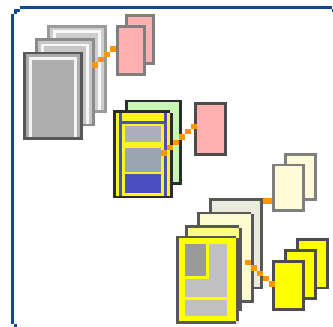
Current Web Pages



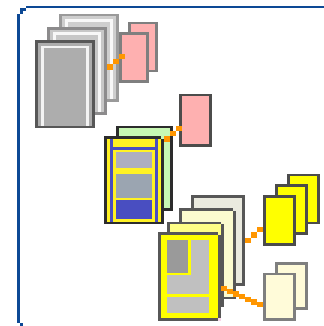
Consolidation,  
Clustering,  
Duplication 제거



와이즈넷 검색 결과



Correct  
Ranking



Structure Results

# 검색엔진의 발전과정

---

- 검색 모델
  - Boolean model → Vector model
- 질의어
  - Keyword 검색 → 자연어 검색
- 부가기능
  - Image, sound
  - 검색결과 clustering
- H/W
  - Workstation → PC server

# 검색엔진 평가 방법

---

- 재현율(recall ratio)
  - 정답문서를 검색한 비율
- 정확률(precision ratio)
  - 검색된 문서 중 정답문서의 비율
- F-measure
  - 재현율과 정확률을 하나의 값으로 표현
- R precision(precision at rank  $n$ )
  - 상위  $n$ 개의 검색결과에 대한 적합한 문서 비율



# “princess diana”의 검색결과

## Engine 1

### Princess Diana Memorial WebRing

Follow the WebRing for a tour of memorial site

87% <http://www.geocities.com/RainForest/Vines/1009/diana>  
1998

Grouped results from <http://www.geocities.com>

### FOR DIANA, PRINCESS OF HEART - Dr. K

...

Dr. Kate Wachs Comments on Princess Diana T

84% <http://www.therelationshipcenter.com/diana.shtml> (S

### Princess Diana Editorial Cartoons! Cartoons :

The Professional Cartoonists Index is the most c  
cartoonists o

daily cartoon  
82% <http://www>

**Relevant and  
high quality**

### Diana, Princess of Wales

1 July 1961 - 31 August 1997 The BBC Web sit  
Camera Press/Snowdon

79% <http://www.royal.gov.uk/start.htm> (Size 2.3K) Doc

Grouped results from <http://www.royal.gov.uk>

## Engine 2

### 1. Re: Lost in the shadow of Princess Diana

[URL: [www.spiceisle.com/talkshop/messages/6232.htm](http://www.spiceisle.com/talkshop/messages/6232.htm)]  
The SpiceIslander TalkShop. [ Follow Ups ] [ Pos  
The SpiceIslander TalkShop ] Date: September  
00:54:03 From: Sno,...  
Last modified 12-Sep-97 - page size 4K - in English [ Tran

### 2. Re: Princess Diana's gown auction

[URL: [www.elle.com/textes/blablaba/forum/messages1/15](http://www.elle.com/textes/blablaba/forum/messages1/15)]  
Re: Princess Diana's gown auction. [ Follow Ups  
Followup ] [ Elle International - Blablaba ] Posted  
September 07, 1997 at 02:15:26:..  
Last modified 30-Mar-98 - page size 2K - in English [ Tran

### 3. Re: Princess Diana

[URL: [spicyhot.com/gaynet/messages/1053.html](http://spicyhot.com/gaynet/messages/1053.html)]  
Re: Prince  
Maine Ga  
Novembe  
Last modifi

**Relevant but  
low quality**

### 4. Re: Princess Diana - Queen of Hearts

[URL: [www.elle.com/textes/blablaba/forum/messages1/26](http://www.elle.com/textes/blablaba/forum/messages1/26)]  
Re: Princess Diana - Queen of Hearts. [ Follow U  
Followup ] [ Elle International - Blablaba ] Posted  
on August 31, 1997 at..  
Last modified 30-Mar-98 - page size 4K - in English [ Tran

## Engine 3

### 1. Free Passwords To Adult Sites ...

99% - Articles & General info: Free Passwords  
Sites ..... warez princess diana demi moore  
magazine kathy ireland lingerie jennifer aniston cook  
warez princess diana demi moore... 03/09/98  
Commercial site: <http://www.purient.com/warez>

### 2. SEX CHAT XXX NUDE PORNO PLAYBOY P

AMERICAN FIRST FREE PICTURAL WOMEN  
99% - Articles & General info: SEX CHAT X  
PORNO PLAYBOY PAMELA AMERSON P  
PICTURES WOMEN ADULT MUSIC CHAT B  
EROTICA JENNIFER MCCARTHY LINGERIE SA  
CHAT CRAWFORD FREE GILLS... 03/09/98

Personal page: <http://www.connix.com/~wggonzo/sex/slidesuperall.htm>

### 3. Ro

with pussy Fucked Lame as Quanta was getting pr  
Personal page: <http://www.octet.com/~gonzo/jy>

### 4. Sunday, 18-Jan-98

99% - Articles & General info: Sunday, 18-Jan-  
CHAT XXX NUDE PORNO PLAYBOY PAME

**Not relevant  
index pollution**

# 인터넷 검색엔진의 성능

---

- 웹 문서 개수
  - 99년12월 약 10억개 → 현재는?
    - 블로그 활성화 등으로 인해 기하급수적으로 증가
  - Web spider(crawler)
- 검색결과 ranking
  - 상위 20~30개 내에 적합한 문서 개수
  - Ranking algorithm
- 웹 문서 갱신 주기
  - 뉴스 등 매일 갱신되어야 하는 것

# 웹의 특성

---

- 안정성 문제
  - 23%/day, 38%/week
- 다양한 자료
  - Text, image, sound, script
  - 다양한 언어의 문서
- 중복 문서
  - Syntactic: 약 30%
  - Semantic: ???
- High linkage : 평균 8 links/page

# 질의어 및 사용자 특성

---

- 질의어 특성
  - 평균 2.35 terms
  - 부정확한 질의어
  - 연산자 없는 질의어: 약 80%
- 사용자 특성
  - 사용자 85% -- one screen only
  - 질의어 78% -- 수정 안함
  - Link를 따라감

# 웹 검색엔진 구성요소

---

- Web Spider(crawler)
  - 웹 문서 수집
- Indexer
  - 색인어 추출 및 색인어 저장 구조
- Search interface
  - 질의어 분석 및 검색

# 웹 정보검색 *issues*

---

- 웹 문서 수집
  - Priority : 매일 갱신되는 page?
  - Load balancing : Internal, external
  - Trap avoidance
    - 서버가 죽어 있는 경우
    - Page가 삭제된 경우
- 문서 처리
  - 중복문서 제거
  - 색인어 추출 및 저장 구조
  - Query-independent ranking
  - 문서 분류

# 웹 정보검색 *issues* (계속)

---

- 질의어 처리
  - Query-dependent ranking
  - 중복문서 제거
  - 질의어 수정/확장
  - 검색결과 clustering

# Ranking 문제

---

- Example

- Query-independent ranking
  - 각 문서에 대한 가중치 부여
- Query-dependent ranking
  - 벡터 모델의 cosine measure

- 문서 분석 기법

- Ad-hoc factors: publication, location
- Human annotation

- 문서 연관성 기법

- Query-independent: PageRank, in-degree
- Query-dependent: HITS



# 문서 연관성 기법 (PageRank)

---

- Idea
  - hyperlink information of the Web
- Assumptions
  - Links often connect related pages
  - A link between pages is a recommendation

$$PR(A) = (1 - d) + d \left( \frac{PR(T1)}{C(T1)} + \frac{PR(T2)}{C(T2)} + \dots + \frac{PR(Tn)}{C(Tn)} \right)$$

# PageRank: Query-independent ranking

---

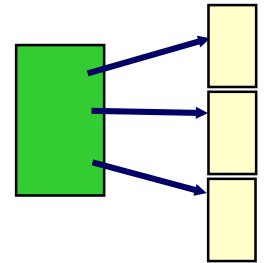
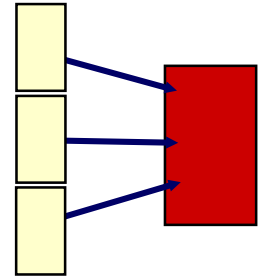
- 웹 페이지의 그래프 표현
  - $(u, v)$  : page  $u$ 에서 page  $v$ 로 link
- 웹 페이지의 quality
  - In-degree 및 그 페이지에 link된 페이지의 quality에 의해 결정
- 웹 페이지의 PageRank는 사용자가 그 페이지에 머무는 시간에 비례
- Google에서 사용하는 ranking 기법 중 하나

# HITS: Query-dependent ranking

---

- Given a query find:

- Good sources of content (**authorities**)
- Good sources of links (**hubs**)



- Better authority comes from** in-edges from **good hubs**. Being a **better hub comes from** out-edges to **good authorities**.

# Modified HITS

---

- 문제점: Some edges are “wrong”
  - Multiple edges from same author
  - Automatically generated
  - 해결방법: Edge weighting
- 문제점: Topic drift
  - 예) jaguar + car → pages about *cars*
  - 해결방법: Analyze content and assign topic scores to nodes

# Connectivity Server

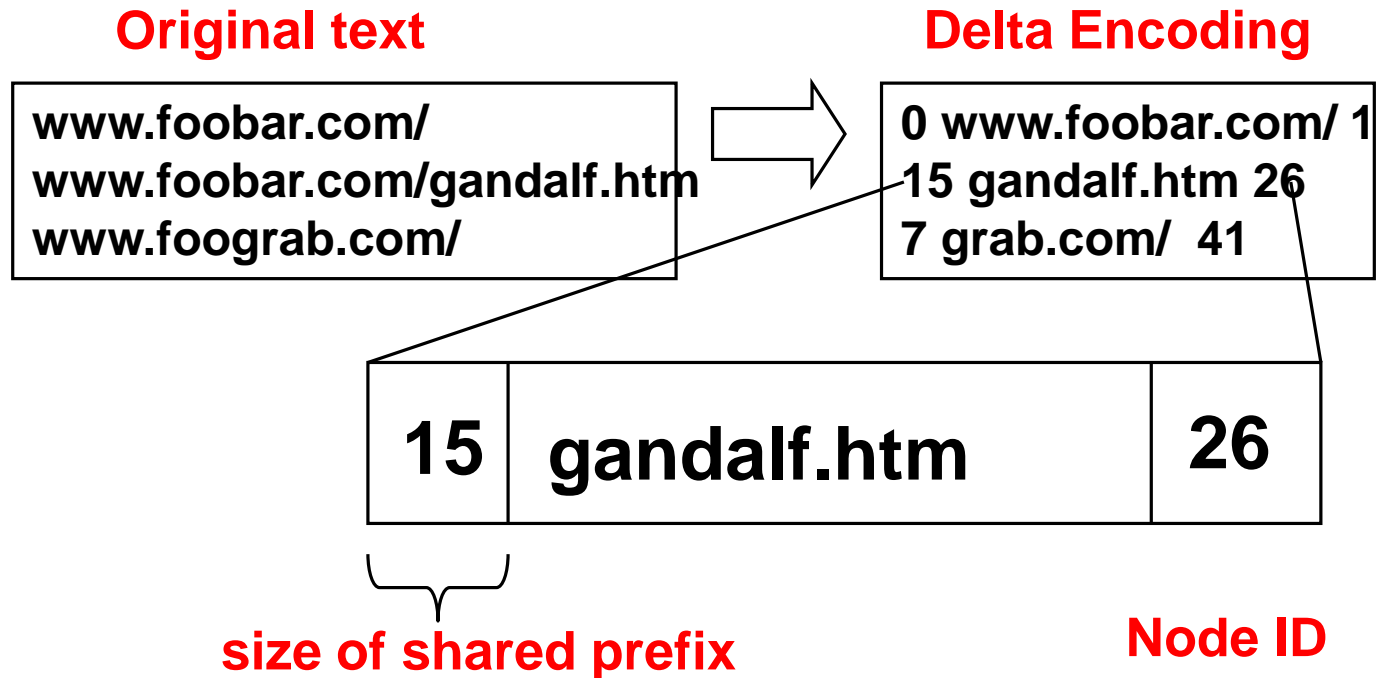
---

- Basic operations
  - InEdges(URL u, int k)
  - OutEdges(URL u, int k)
- Difficulties
  - Memory usage: 180M nodes, 1B edges
  - Preprocessing time: days
  - Query time: 0.0001sec/result URL

# URL database

Sorted list of URLs is 8.7 GB ( $\approx 48$  bytes/URL)

→ Delta encoding reduces it to 3.8 GB ( $\approx 21$  bytes/URL)



# Other I.R. issues

---

- Duplicate filtering : 중복 문서 제거
- 갱신 주기 문제
- 검색결과 관련
  - Clustering → [www.northernlight.com](http://www.northernlight.com)
  - Summarization
- Directory service
  - Document classification
- 분야별 전문화된 검색엔진

# *Duplicate filtering*

---

- Near-duplicate documents
  - Computing pair-wise edit distance
  - A short sketch for each document
- Near-duplicate hosts(mirrors)
  - Pre-filtering techniques
    - IP-based
    - URL-string based
      - Similar hostnames, similar paths
    - URL-string & hyperlink based
    - Hostname & hyperlink based

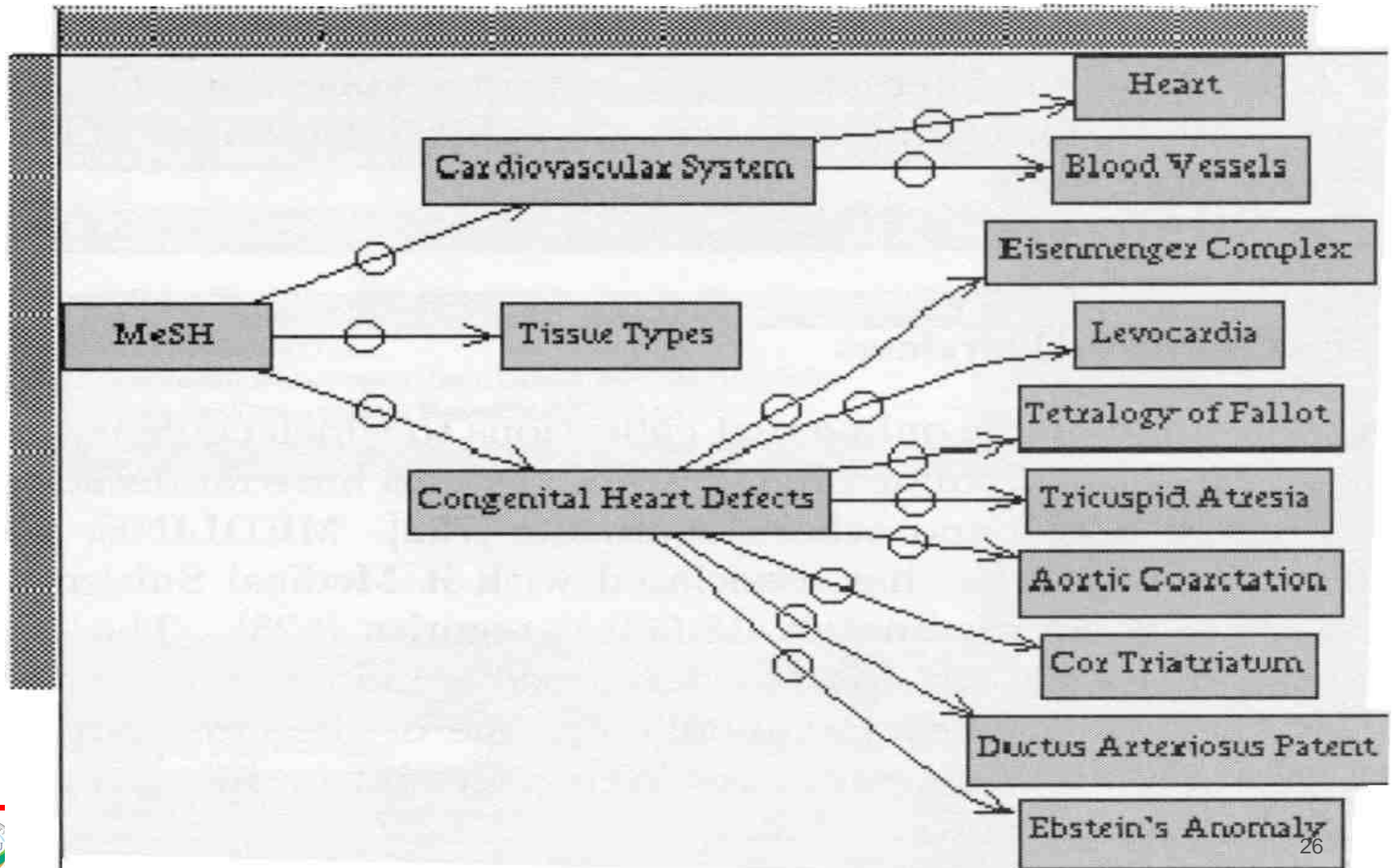


# *User interface & visualization*

---

- Category or directory overview
  - MeSHBrowse
  - Scatter/Gather
  - 2/3-dimensional overview
- Query specification
  - Venn diagram
  - Filter-flow visualization
  - Block-oriented diagram visualization
- Current document set in the context of other information types

# MeSHBrowse interface for category labels



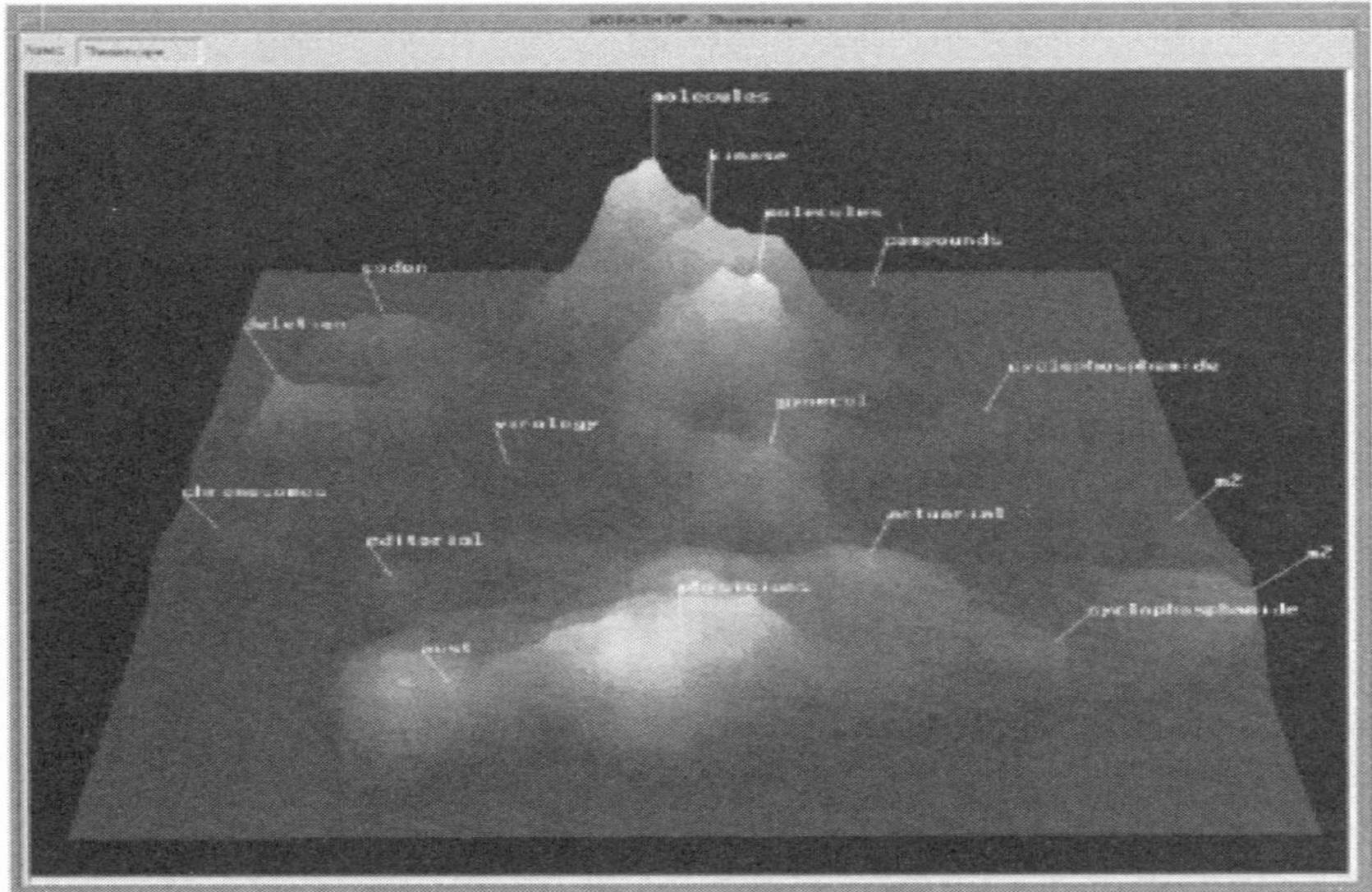
# Scatter/Gather clustering

<input type="checkbox"/> Cluster 1 Size: 8	key army war francis spangle banner air song scott word poem british
<input type="radio"/> Star-Spangled Banner, The	
<input type="radio"/> Key, Francis Scott	
<input type="radio"/> Fort McHenry	
<input type="radio"/> Arnold, Henry Harley	
<input type="radio"/> Mibinok, Anthem	
<input type="checkbox"/> Cluster 2 Size: 68	film play career win television role record award york popular stage p
<input type="radio"/> Burstyn, Ellen	
<input type="radio"/> Stanwyck, Barbara	
<input type="radio"/> Berle, Milton	
<input type="radio"/> Zukor, Adolph	
<input type="radio"/> Deakland, Tallulah	
<input type="checkbox"/> Cluster 3 Size: 97	bright magnitude cluster constellation line type contain period spectr
<input type="radio"/> star	
<input type="radio"/> Galaxy, The	
<input type="radio"/> extragalactic systems	
<input type="radio"/> interstellar matter	
<input type="radio"/> cluster, star	



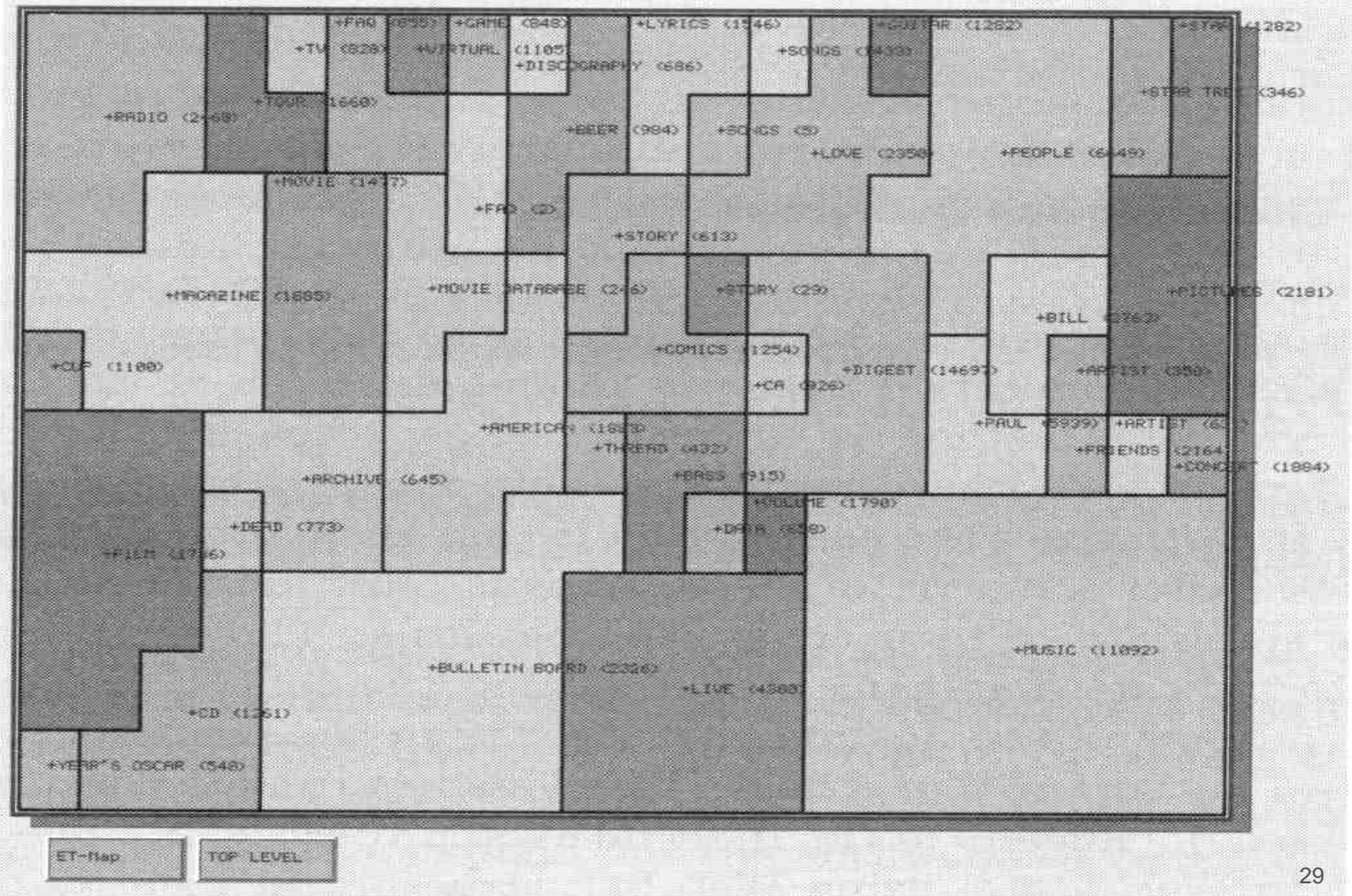
# Three-dim. clustering

---



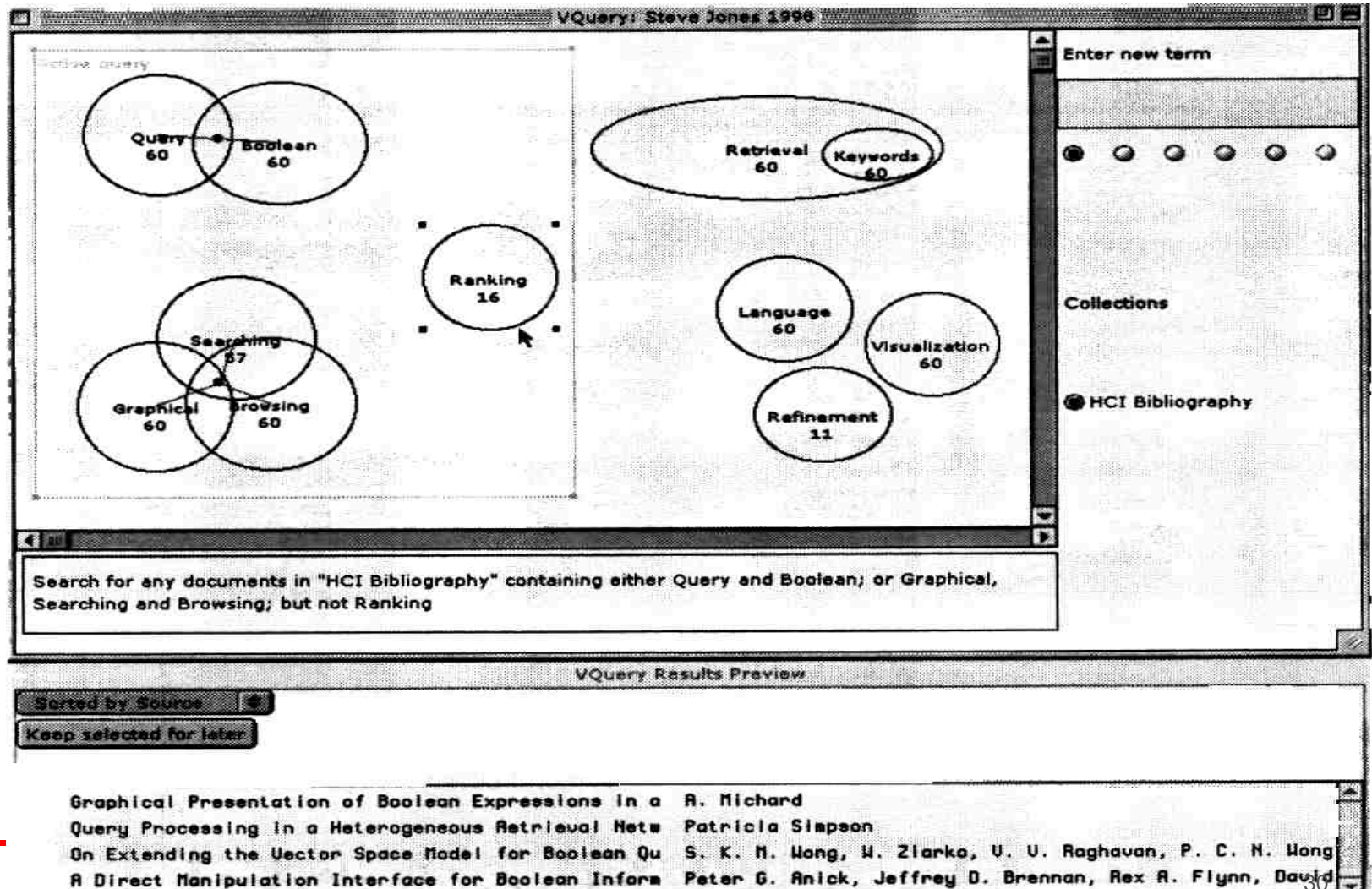


# Two-dim. Web pages

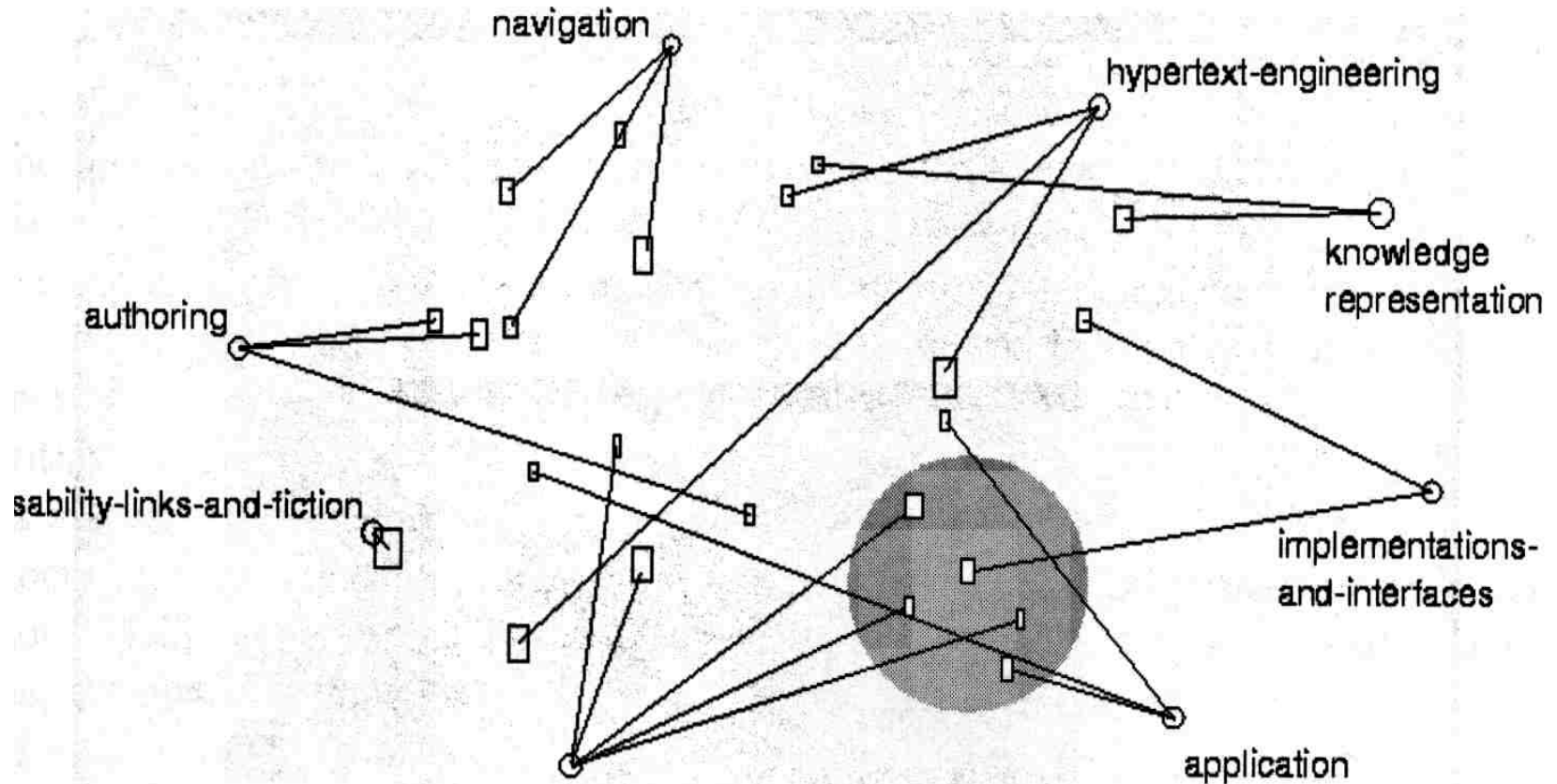




# Venn diagram visualization



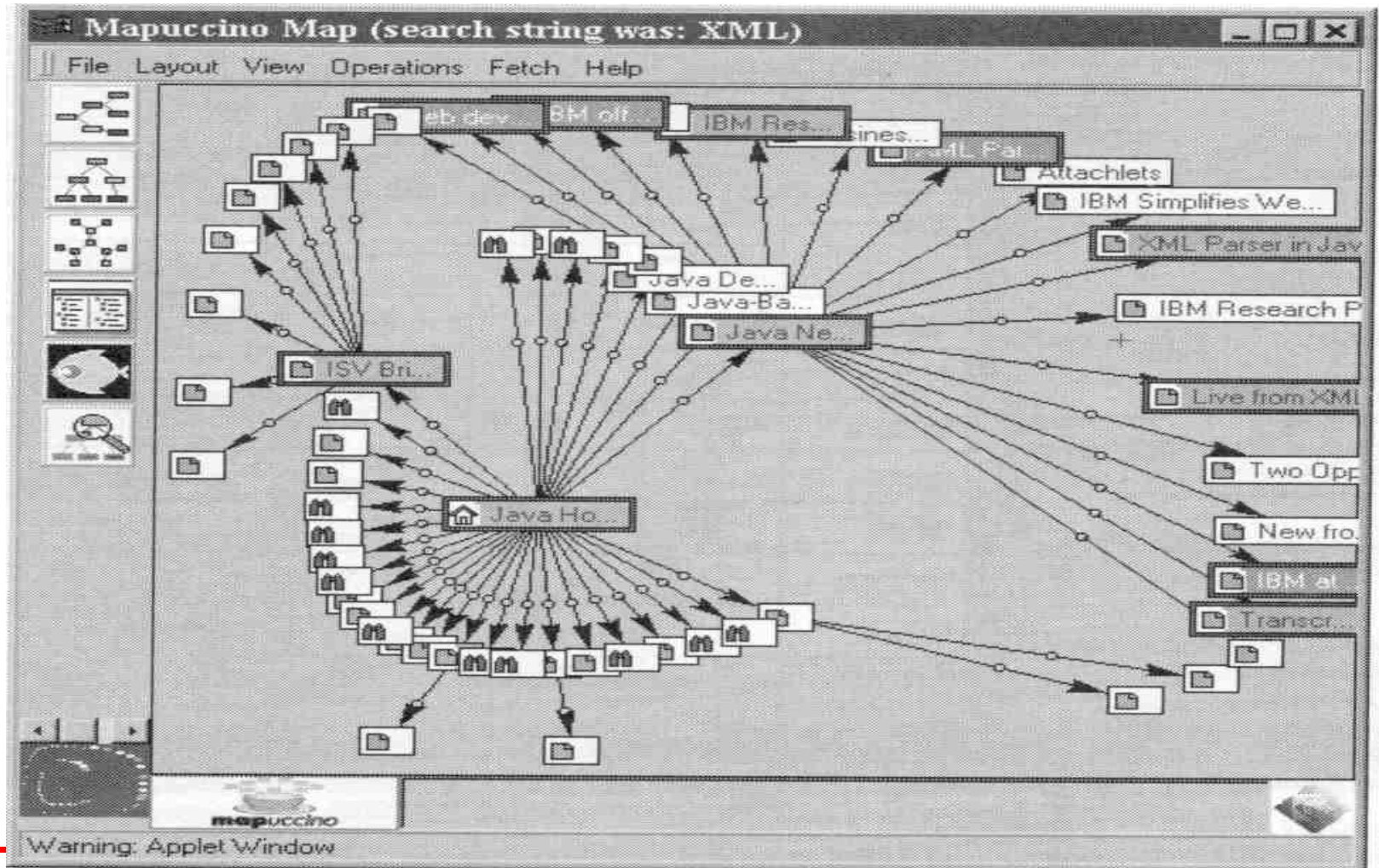
# Query-terms placed in abstract graphical space



**Figure 10.18** An example of the VIBE retrieval results display [452].



# Graphical depiction of Web link structure





- 
- 웹 검색엔진 소개
    - 웹 문서 특성 및 사용자 특성 고찰
  - 웹 정보검색 issues
    - 문서 수집
    - Ranking: 문서 연관성 기법
    - PageRank, HITS
  - <참고> Monika Henzinger의 Google 자료
    - “Web Information Retrieval”