

Dynamic Extraction of Topic Descriptors and Discriminators: Towards Automatic Context-Based Topic Search

Ana Maguitman, David Leake,

Thomas Reichherzer and Filippo Menczer

Indiana University



Joint project with IHMC

Partially supported by NASA under award No NCC 2-1216

Alternative Views of Knowledge Management

■ Knowledge acquisition:

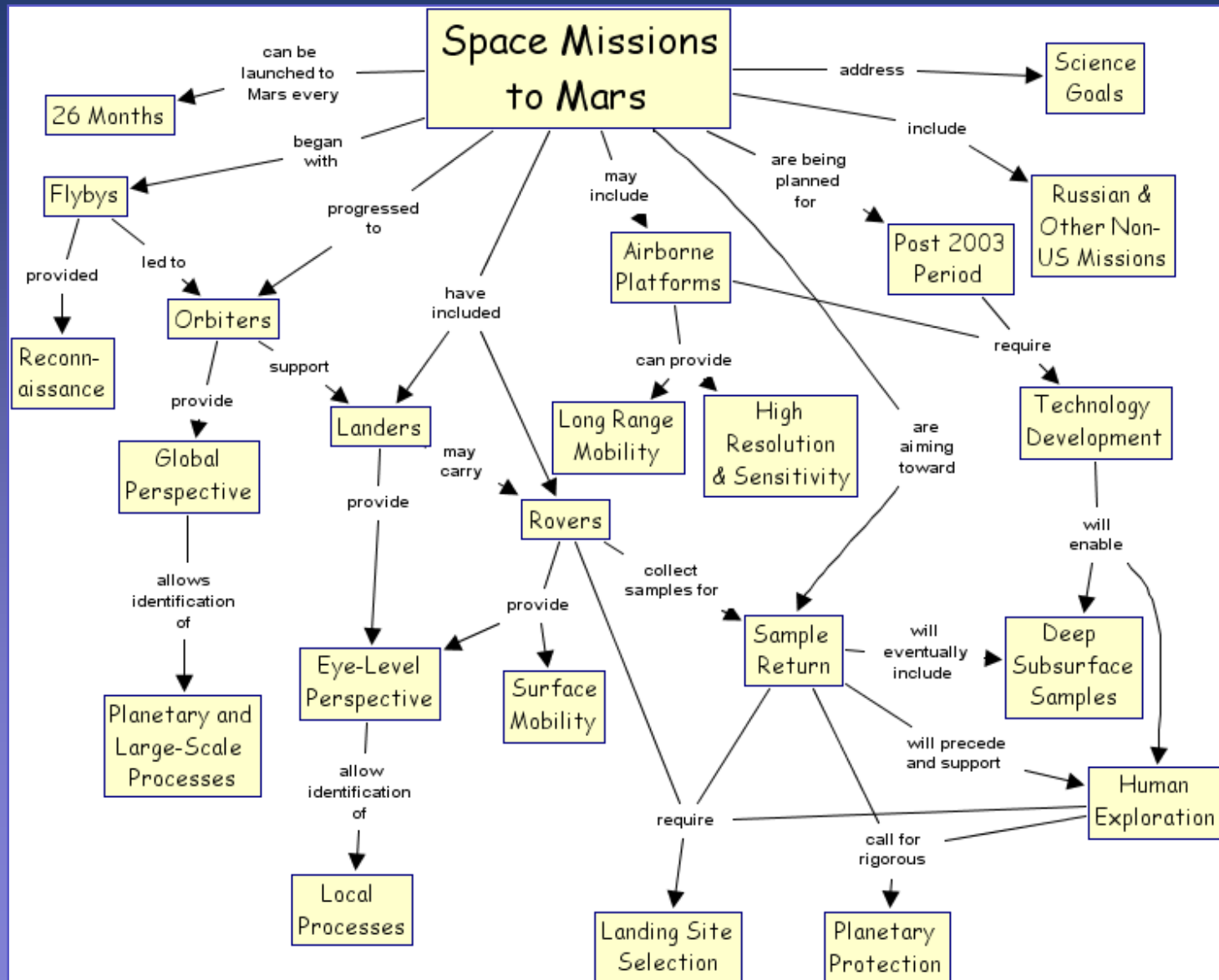
Capture knowledge that is already within the expert.

■ Knowledge extension:

Knowledge models evolve from coordinated processes of knowledge acquisition and knowledge construction.



Concept Maps [Novak, 1977]

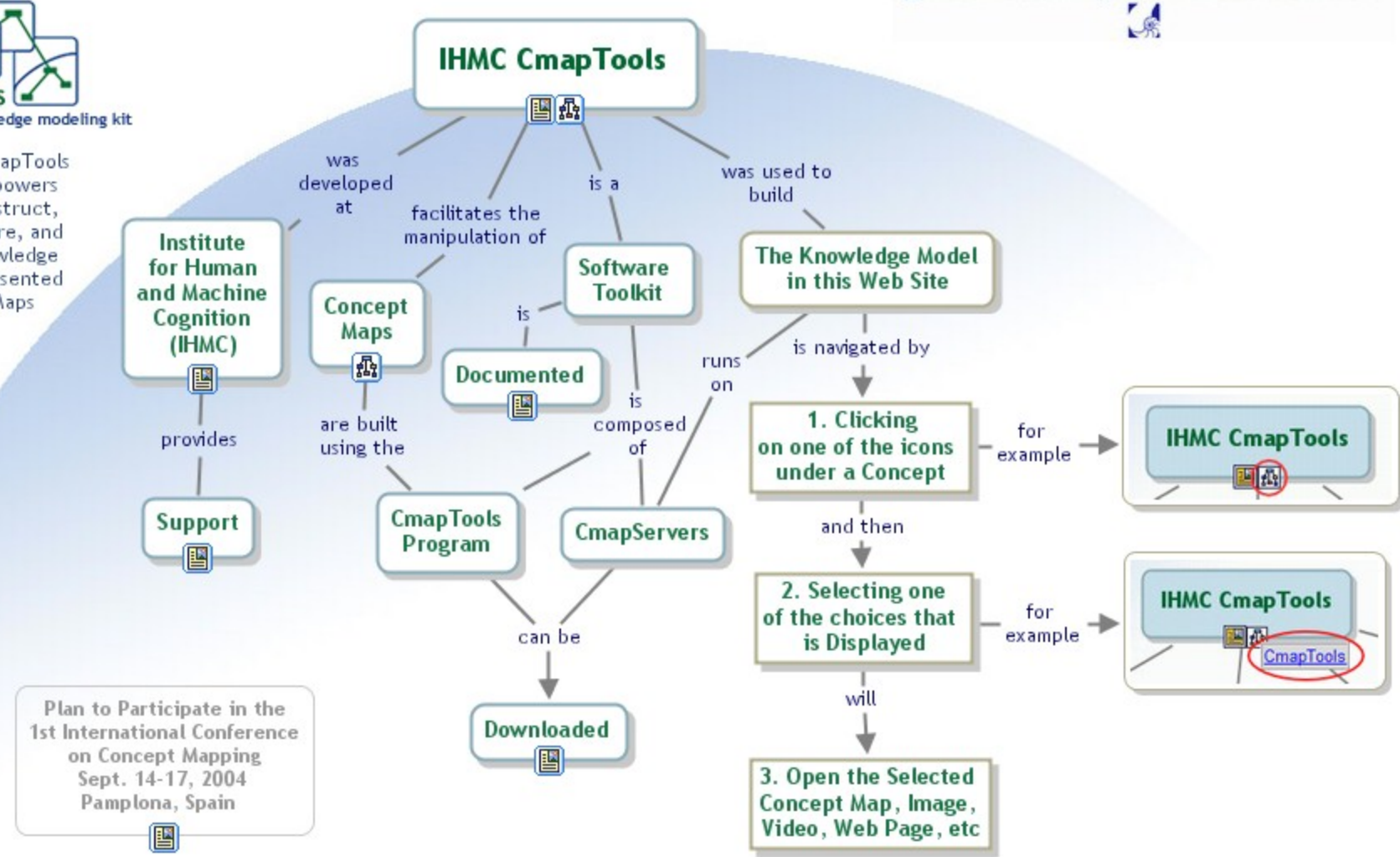




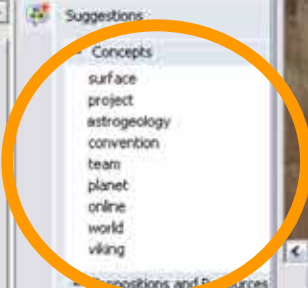
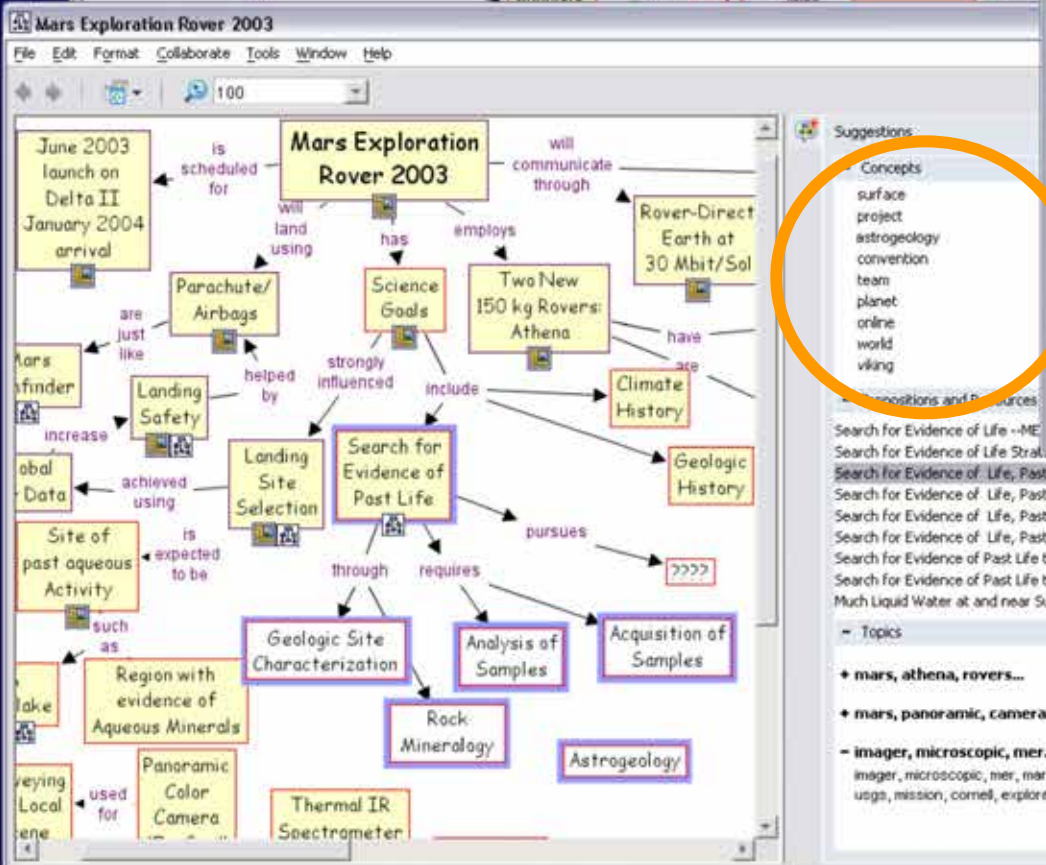
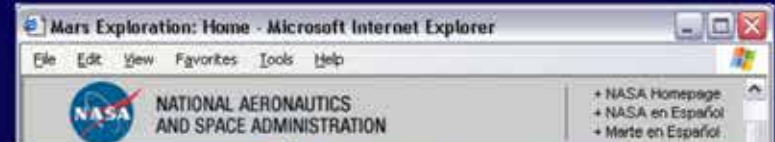
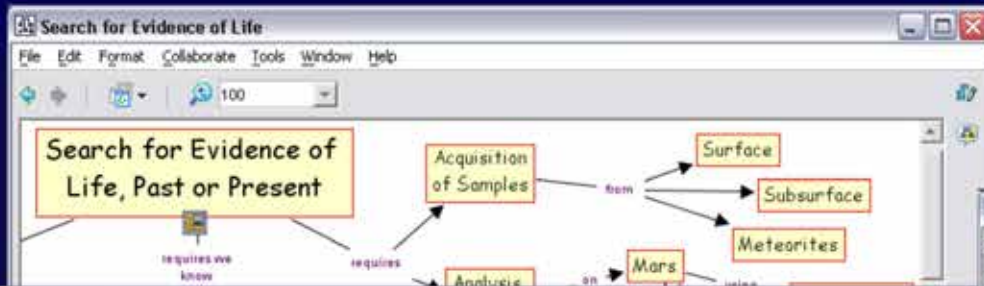
The IHMC CmapTools software empowers users to construct, navigate, share, and criticize knowledge models represented as Concept Maps

Welcome to the Web Site of

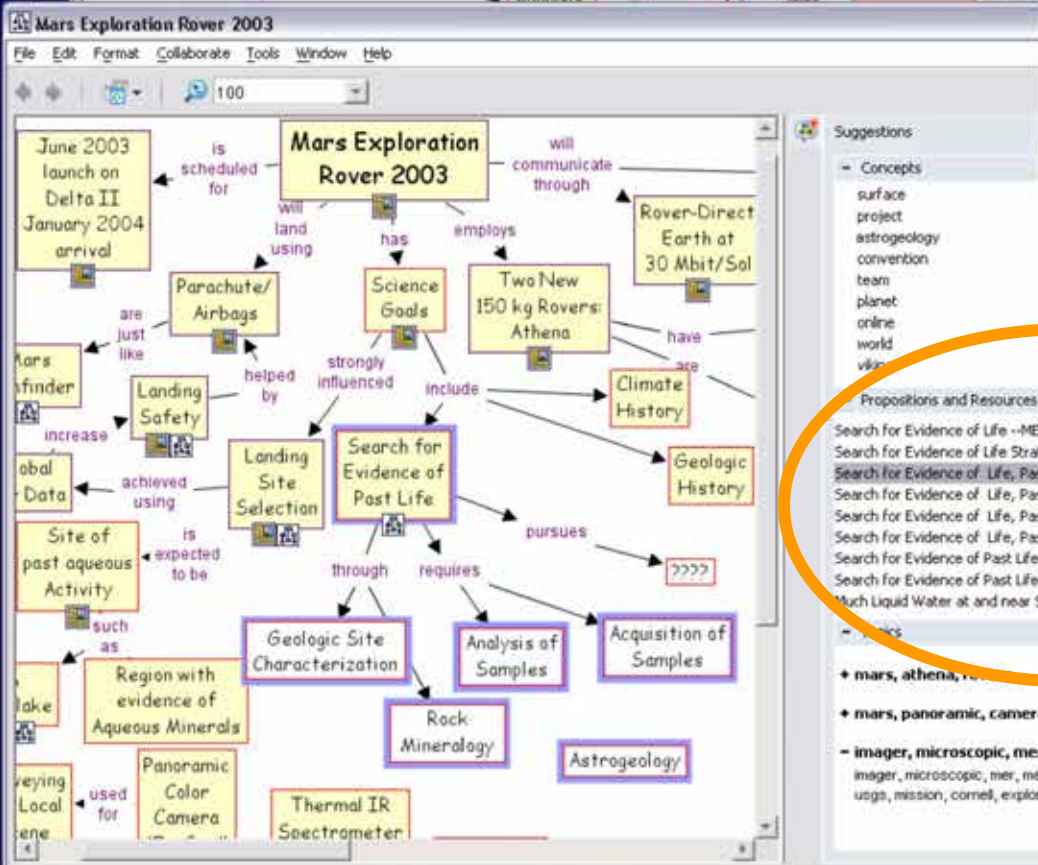
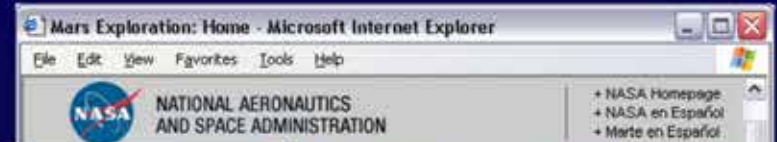
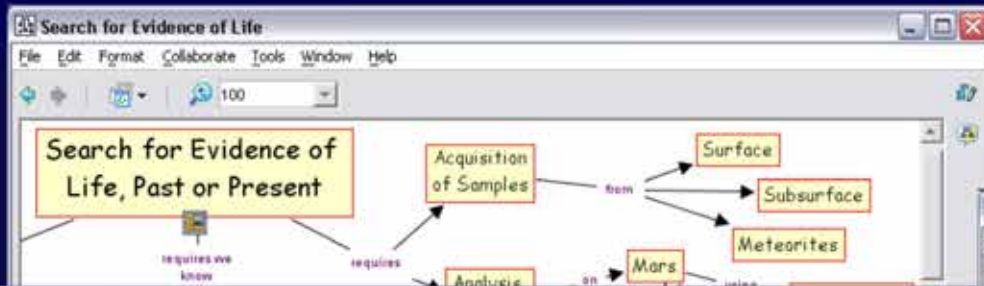
IHMC - A University Affiliated Research Institute



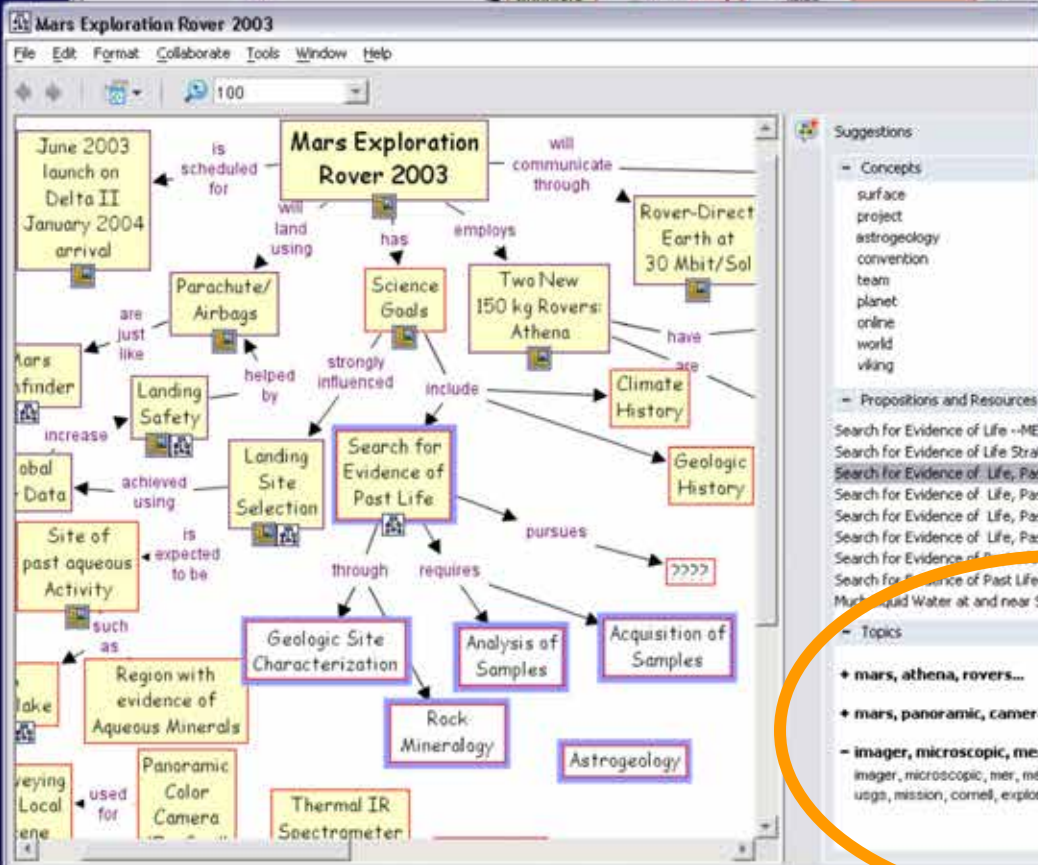
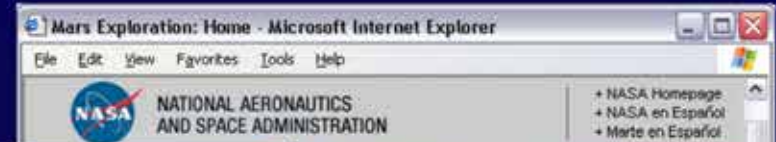
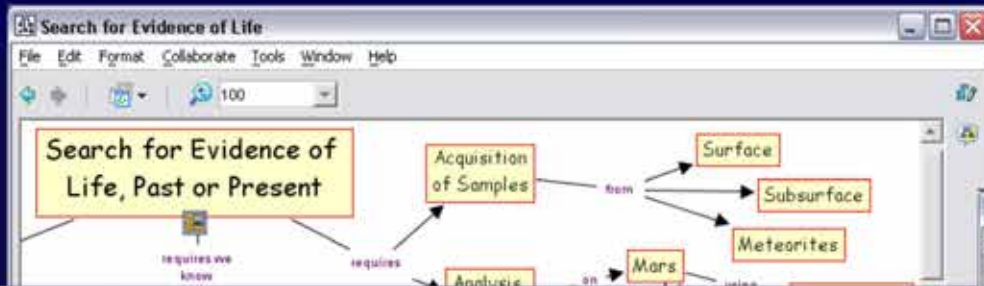
A Family of Suggesters for CmapTools



A Family of Suggesters for CmapTools



A Family of Suggesters for CmapTools



- Suggestions:
- Concepts
 - surface
 - project
 - astrogeology
 - convention
 - team
 - planet
 - online
 - world
 - viking
 - Propositions and Resources
 - Search for Evidence of Life --ME
 - Search for Evidence of Life Strat...
 - Search for Evidence of Life, Past or Present pursues a goal of ASTROBIOLOG...
 - Search for Evidence of Life, Past or Present requires Acquisition of Samples
 - Search for Evidence of Life, Past or Present requires we know WHAT to See
 - Search for Evidence of Life, Past or Present requires Analysis of Samples
 - Search for Evidence of Life, Past or Present requires Geologic Site Characterization
 - Search for Evidence of Life, Past or Present requires Rock Mineralogy
 - Search for Evidence of Life, Past or Present requires Thermal IR Spectrometer
 - Search for Evidence of Life, Past or Present requires Panoramic Color Camera
 - Much Liquid Water at and near Surface in the Past motivates Search for Evid...
 - Topics
 - + mars, athena, rovers...
 - + mars, panoramic, camera...
 - imager, microscopic, mer...
 - imager, microscopic, mer, mars, astrogeology, camera, athena, rover, usgs, mission, cornell, exploration, rovers, page, project, home, nasa



Goals for a Topic Suggester

- Make suggestions at a *higher level of topics* (rather than individual resources.)
- Satisfy criteria for topic quality:
 - Novelty
 - Diversity
 - Global Coherence
 - Coverage
 - Local quality



Road Map for Topic Suggestions

- Theoretical framework.
- Application of the theory in the implementation of a topic suggester system.
- Experimental Study.



Term Importance

- n Importance of given term depends on task.
- n Traditional views distinguish two notions:
 - Descriptors: terms that occur frequently in a document.
 - Discriminators: terms that occur in a document but rarely occur in the corpus.



New Challenges for Formulation of Descriptors and Discriminators

- Search methods that can reflect extensive contextual information.
- Methods for topic search.
- Methods for searching open collections of documents.



Topics, Terms and Documents

Mars
Exploration
Surveyor
Global
Opportunity
Spirit
Orbiter
Site
Landing
Camera
Rover
Selection
Images
MGS
MOC



Topics, Terms and Documents

Mars
Exploration
Surveyor
Global Opportunity Spirit
Orbiter Site
Camera Landing
Rover
Selection MGS MOC
Images



Topics, Terms and Documents

Mars
Exploration
Surveyor
Global
Opportunity
Spirit
Orbiter
Site
Landing
Camera
Rover
Selection
Images
MGS
MOC



Topic Descriptors and Discriminators

- Terms are *good topic descriptors* if they answer the question “What is this topic about?”
- Terms are *good topic discriminators* if they answer to the question “What are good query terms to access similar information?”

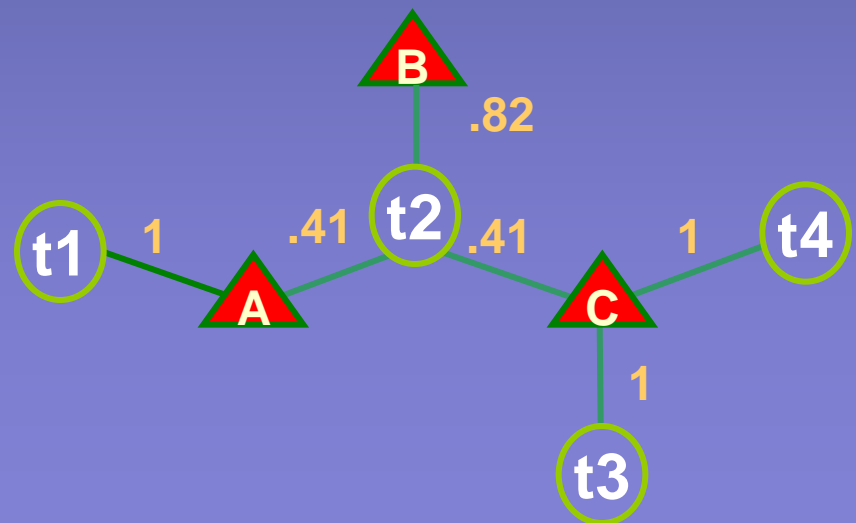
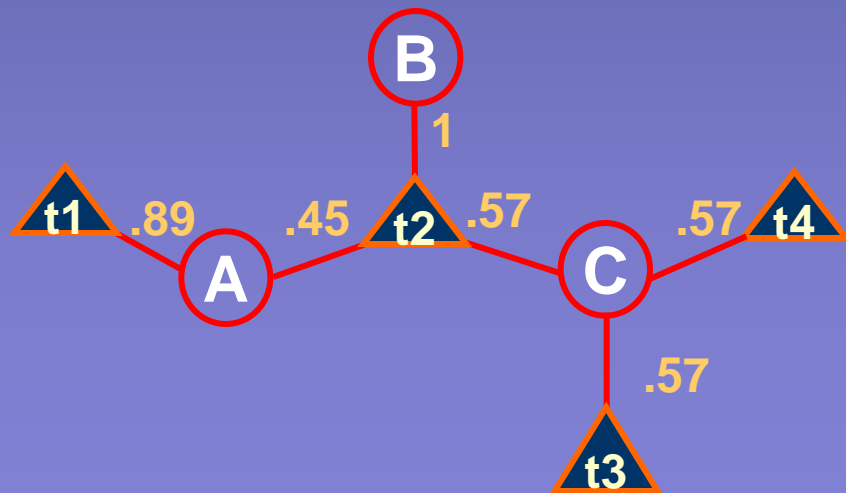
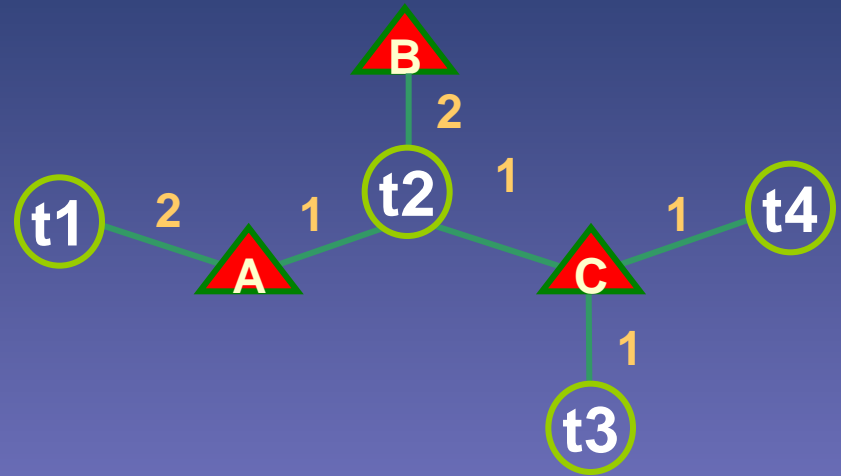
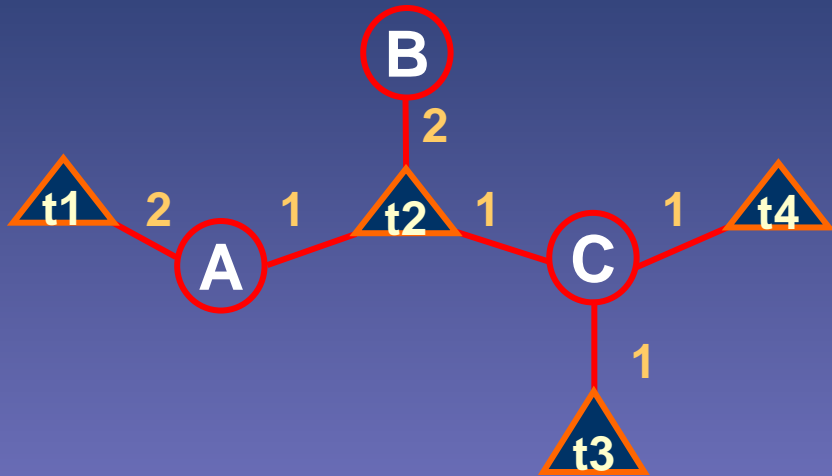


Hypotheses

- I. Terms that tend to occur frequently in the context of a given topic are good topic descriptors.
- II. Terms that tend to occur only in the context of a topic are good topic discriminators.



Using Hypergraph Representations for Documents and Terms



Document-Term Duality

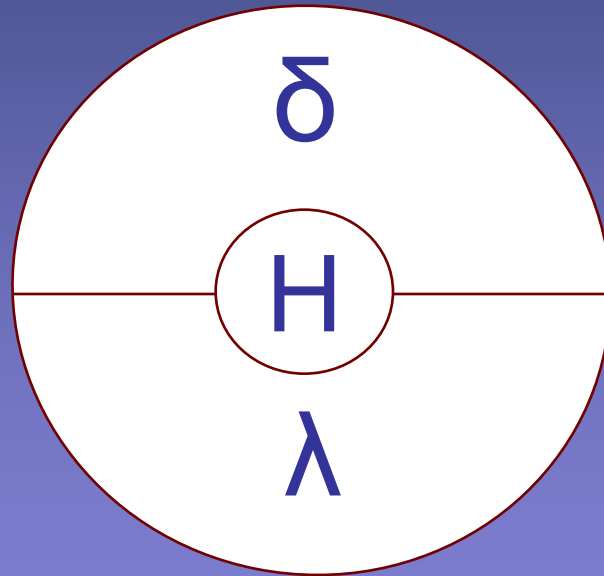


$H[i,j]$: number of occurrences of t_j in d_i

Document-Term Duality

$$\delta(t_i, d_j) = \frac{\text{sgn}(\mathbf{H}^T[i, j])}{\sqrt{\sum_{k=0}^{m-1} \text{sgn}(\mathbf{H}^T[i, k])}}$$

discriminating power of t_i in d_j



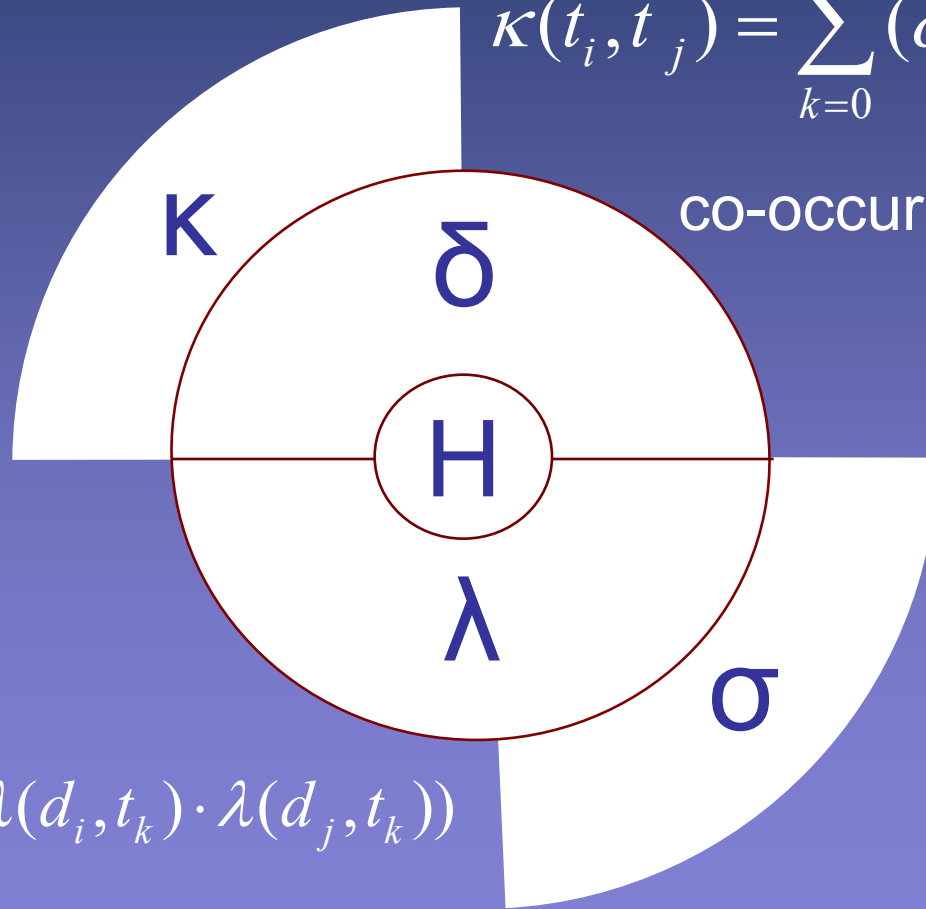
$$\lambda(d_i, t_j) = \frac{\mathbf{H}[i, j]}{\sqrt{\sum_{k=0}^{n-1} (\mathbf{H}[i, k])^2}}$$

descriptive power of t_j in d_i

Document-Term Duality

$$\kappa(t_i, t_j) = \sum_{k=0}^{m-1} (\delta(t_i, d_k) \cdot \delta(t_j, d_k))$$

co-occurrence of t_i and t_j

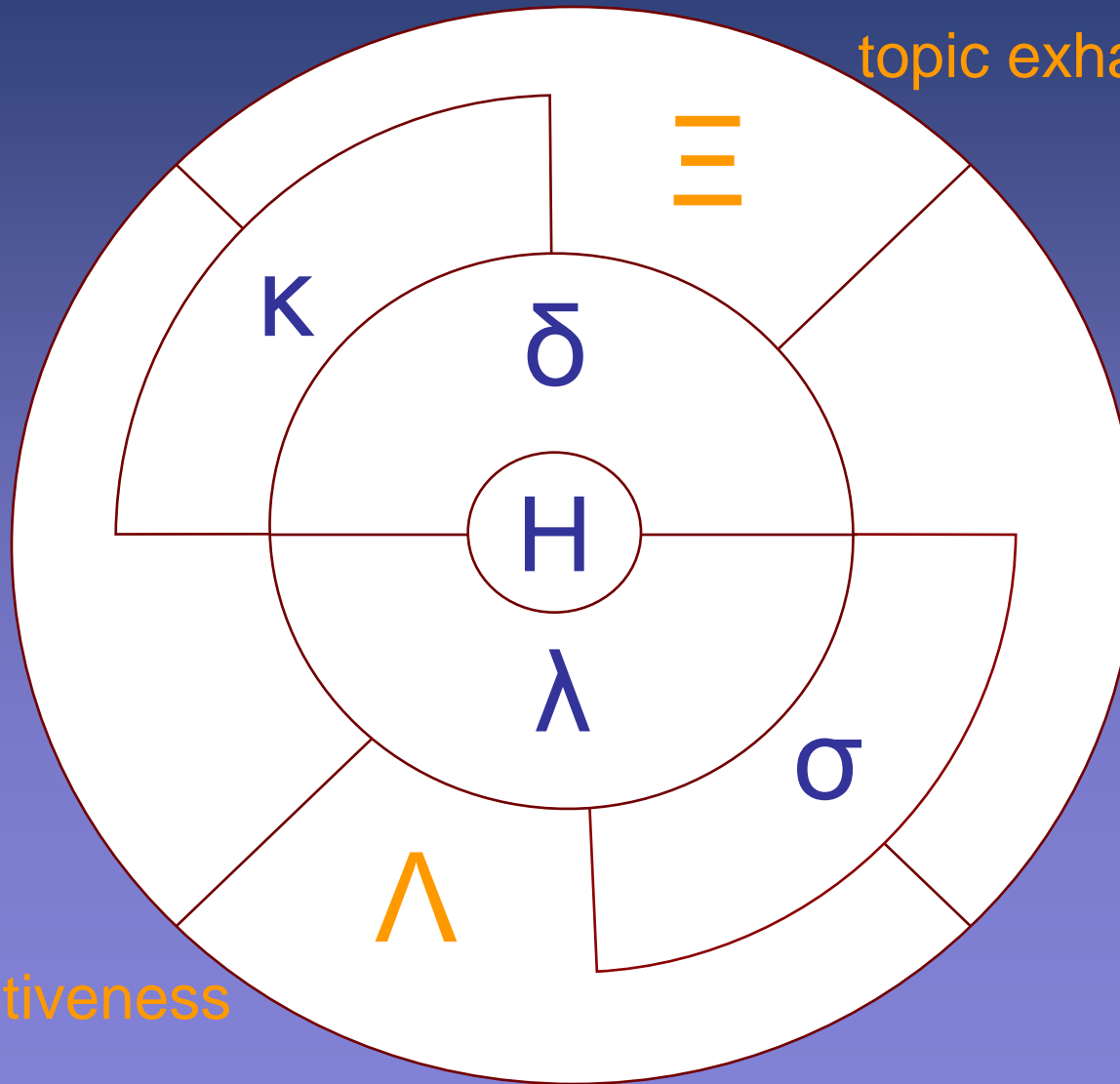


$$\sigma(d_i, d_j) = \sum_{k=0}^{n-1} (\lambda(d_i, t_k) \cdot \lambda(d_j, t_k))$$

similarity between d_i and d_j

Document-Term Duality

topic exhaustivity



topic
descriptiveness

Document-Term Duality

Topic Descriptiveness

$$\Lambda(d_i, t_j) = \begin{cases} 0 & \text{if } \sum_{k=0, k \neq i}^{m-1} \sigma(d_i, d_k) = 0 \\ \frac{\sum_{k=0, k \neq i}^{m-1} (\sigma(d_i, d_k) \cdot \lambda(d_k, t_j)^2)}{\sum_{k=0, k \neq i}^{m-1} \sigma(d_i, d_k)} & \text{otherwise} \end{cases}$$

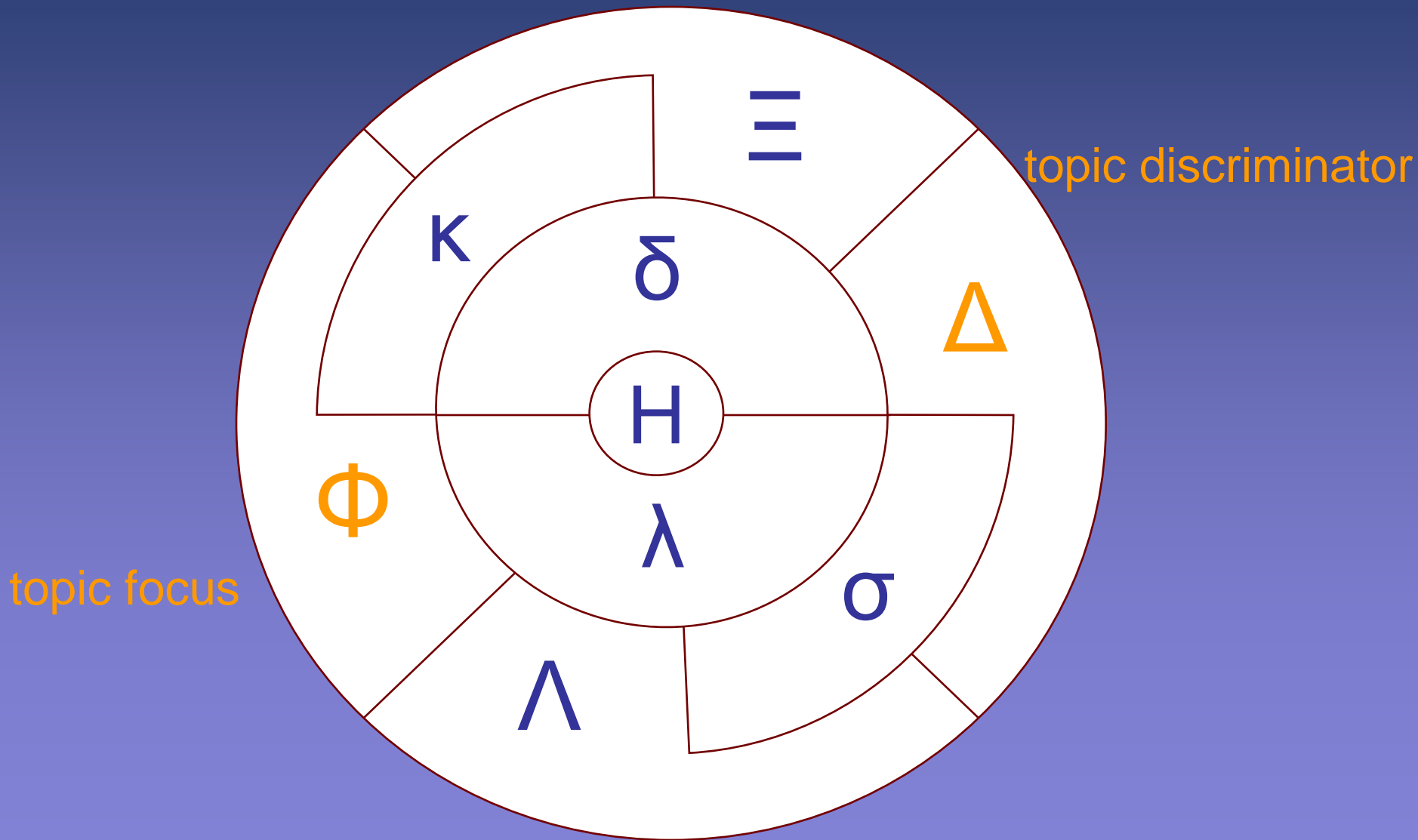
Most documents similar to d_i tend to be described by t_j

Topic Exhaustivity

$$\Xi(t_i, d_j) = \begin{cases} 0 & \text{if } \sum_{k=0, k \neq i}^{n-1} \kappa(t_i, t_k) = 0 \\ \frac{\sum_{k=0, k \neq i}^{n-1} (\kappa(t_i, t_k) \cdot \delta(t_k, d_j)^2)}{\sum_{k=0, k \neq i}^{n-1} \kappa(t_i, t_k)} & \text{otherwise} \end{cases}$$

Most terms that co-occur with t_i tend to discriminate d_j

Document-Term Duality



Document-Term Duality

Topic Discriminator

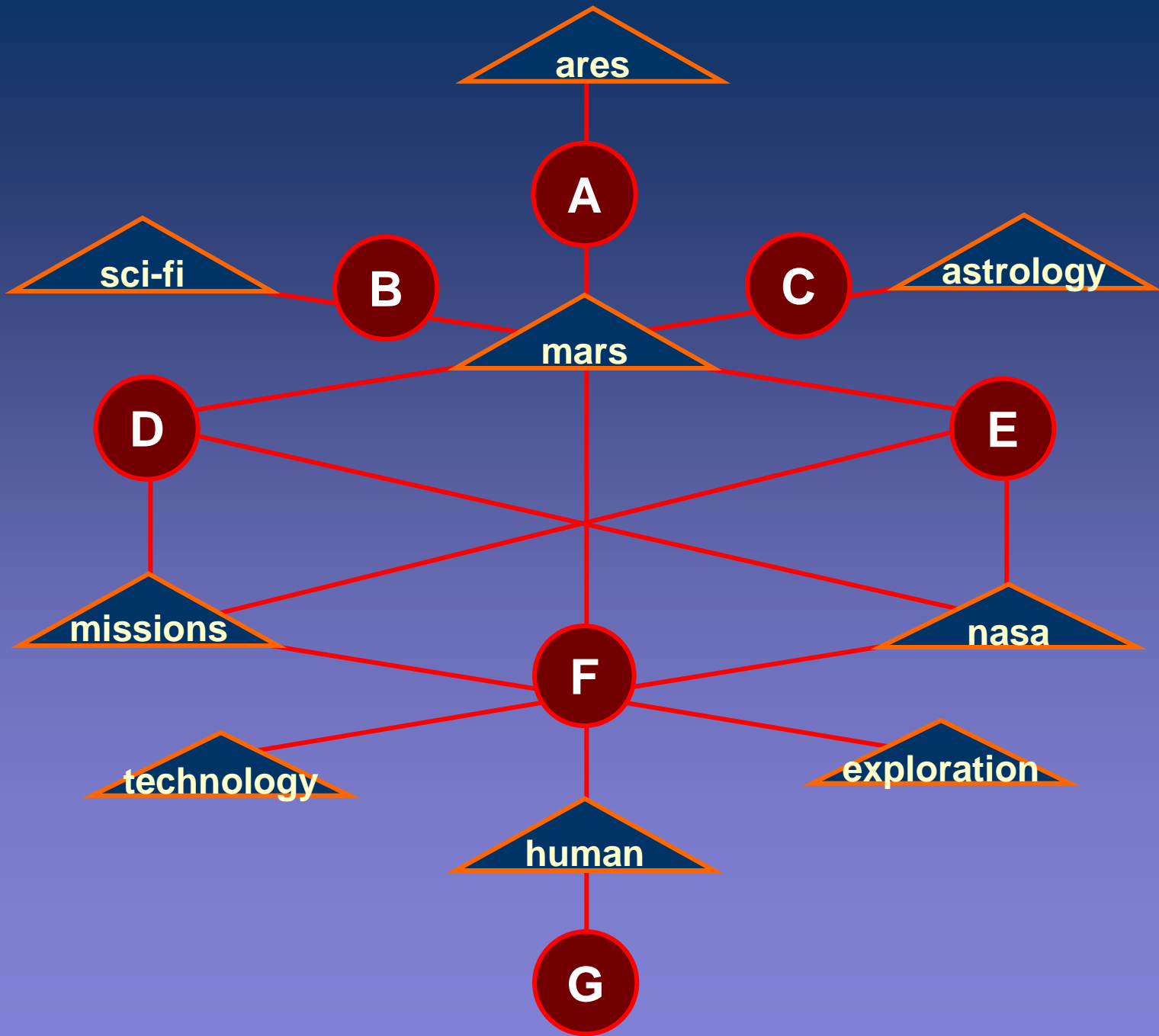
$$\Delta(t_i, d_j) = \sum_{k=0, k \neq j}^{m-1} (\sigma(d_k, d_j) \cdot \hat{\partial}(t_i, d_k)^2)$$

Documents discriminated by t_i are similar to d_j

Topic Focus

$$\Phi(d_i, t_j) = \sum_{k=0, k \neq j}^{n-1} (\kappa(t_k, t_j) \cdot \lambda(d_i, t_k)^2)$$

Terms describing d_i tend to co-occur with t_j



Applying the Theory

- Descriptors and Discriminators are used as query terms to favor recall and precision.
- Similarity is used to filter irrelevant documents.
- The higher order dual notions are applied in the implementation of a co-clustering algorithm to obtain cohesive topics.
- Descriptors are good topic labels.
- A combination of focus and exhaustivity is used to rank documents in a topic.

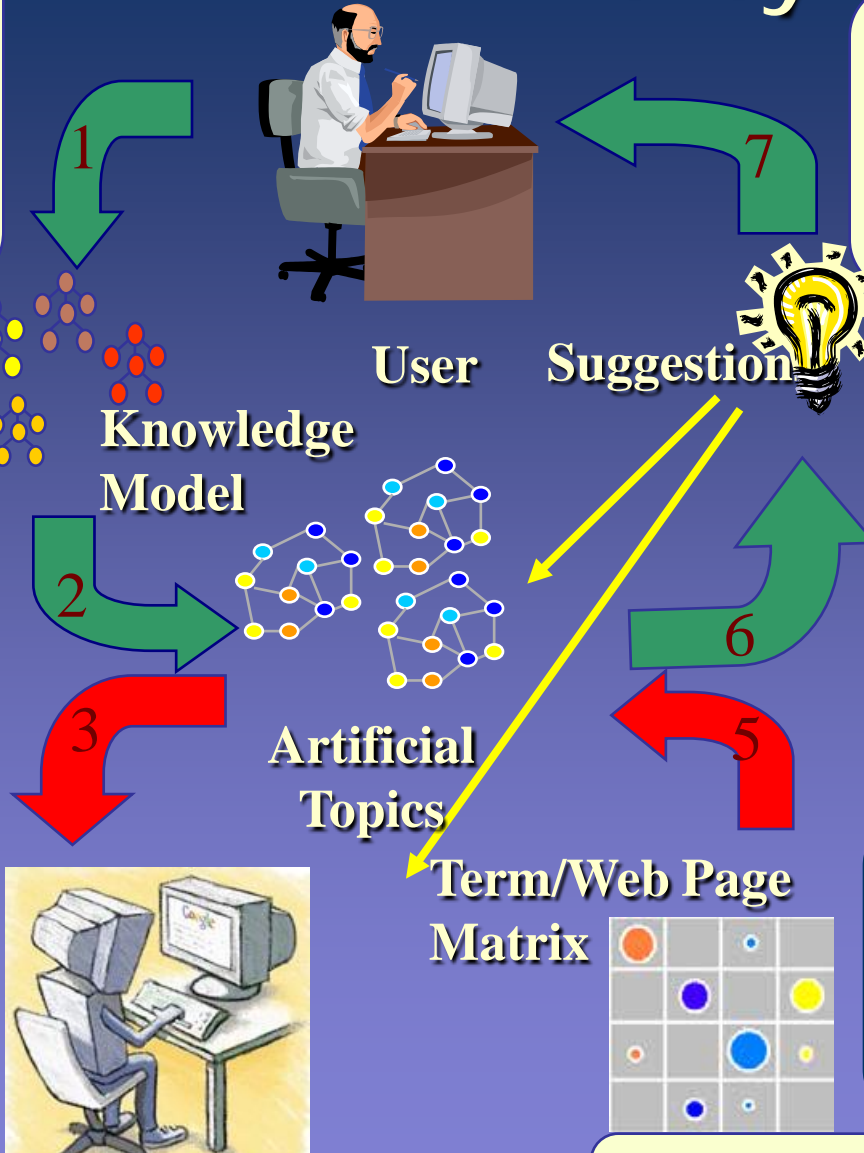


EXTENDER's Cycle

1. EXTENDER starts with a Knowledge Model under construction

2. Combines terms found in concept maps to produce the first generation of Topics

3. Incrementally searches the Web for novel but relevant material



7. Novel topics are presented as suggestions to the user

Novel Terms
Propositions

6. Characterizes novel topics as sets of terms and relevant Web pages

5. Applies clustering techniques to return the next generation of Topics

4. Fills in term/Web page matrix

Google Web API

4

Extender Suggesting New Topics

The image displays a web browser window with a central topic map and a sidebar of suggestions. The topic map is a network of interconnected concepts related to Mars exploration, including:

- Mars Exploration Rover 2003** (central node)
- Parachute/Airbags** (connected to Mars Exploration Rover 2003 via "will land using")
- Science Goals** (connected to Mars Exploration Rover 2003 via "has")
- Two New 150 kg Rovers: Athena** (connected to Mars Exploration Rover 2003 via "employs")
- Rover-Direct Earth at 30 Mbit/Sol** (connected to Mars Exploration Rover 2003 via "will communicate through")
- Landing Site Selection** (connected to Science Goals via "strongly influenced")
- Geologic History** (connected to Science Goals via "include")
- Climate History** (connected to Science Goals via "include")
- Site of past aqueous Activity** (connected to Landing Site Selection via "is expected to be")
- Geologic Site Characterization** (connected to Landing Site Selection via "achieved using")
- Rock Mineralogy** (connected to Geologic Site Characterization via "will be assessed by")
- Onboard Instruments** (connected to Geologic Site Characterization via "will be assessed by")
- Rock Abrasion Tool** (connected to Onboard Instruments via "will be assessed by")
- Region with evidence of Aqueous Minerals** (connected to Site of past aqueous Activity via "such as")
- Possible 1-year Extended Mission** (connected to Climate History via "will have")
- Deep Space Network** (connected to Earth via "has")
- Antennas** (connected to Deep Space Network via "has")
- Orbiters** (connected to Mars Global Surveyor via "????")
- Odyssey** (connected to Orbiters via "????")
- Opportunity** (connected to Orbiters via "????")
- Gusev Crater** (connected to Mars Global Surveyor via "????")
- Lander** (connected to Mars Global Surveyor via "????")
- robot field geologists** (connected to Mars Global Surveyor via "????")
- landing site** (connected to Mars Global Surveyor via "????")

The sidebar on the right, titled "Suggestions", lists several topic suggestions:

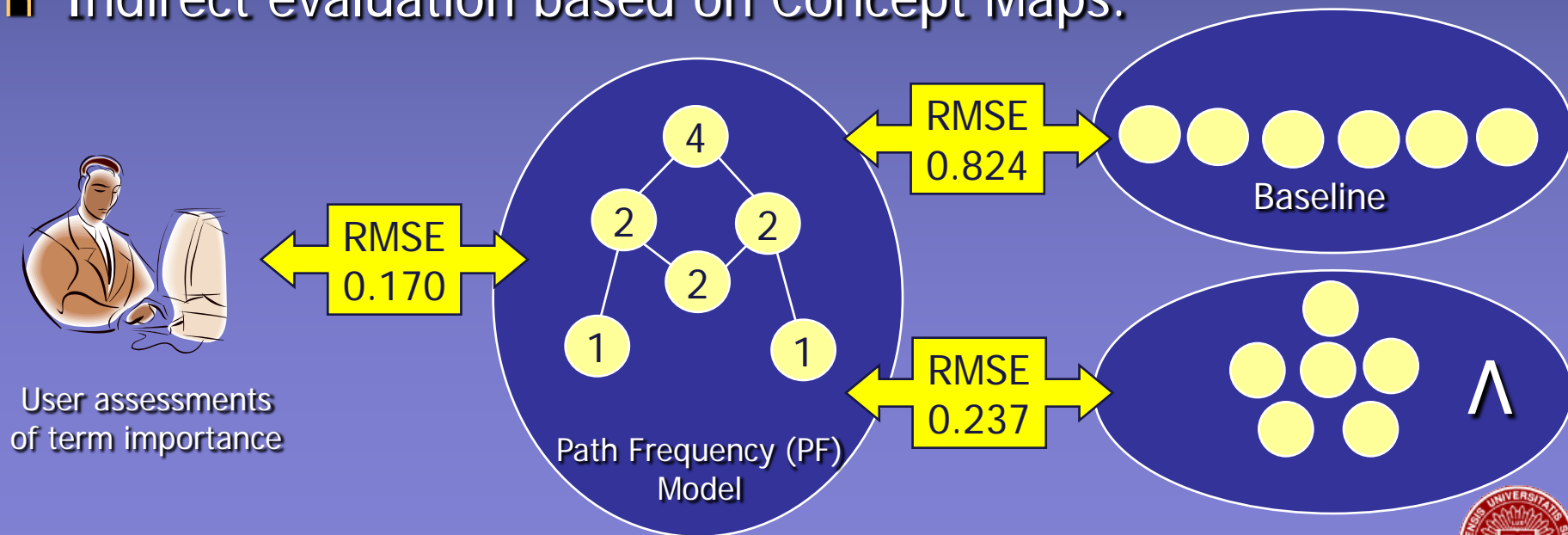
- main, results, mission...
- rover, spirit, mars... (highlighted)
- martian, mars, atmosphere...
- human, nasa, exploration...

The background shows a NASA website with a Mars Explorer satellite image and various navigation links like "Mars for Kids", "Mars for Students", etc.

Evaluation of Descriptor Extraction Method

How good is the descriptor extraction method in finding terms that are good descriptors of a topic?

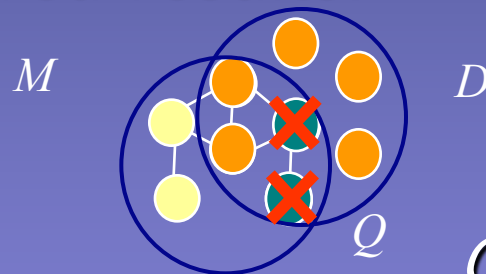
- It is difficult to evaluate descriptive power objectively.
- Indirect evaluation based on Concept Maps.



Evaluation of Discriminator Extraction Method

How good is the query formation process based on the use of topic discriminators?

- Provide an approximate measure of relevance of retrieved results.



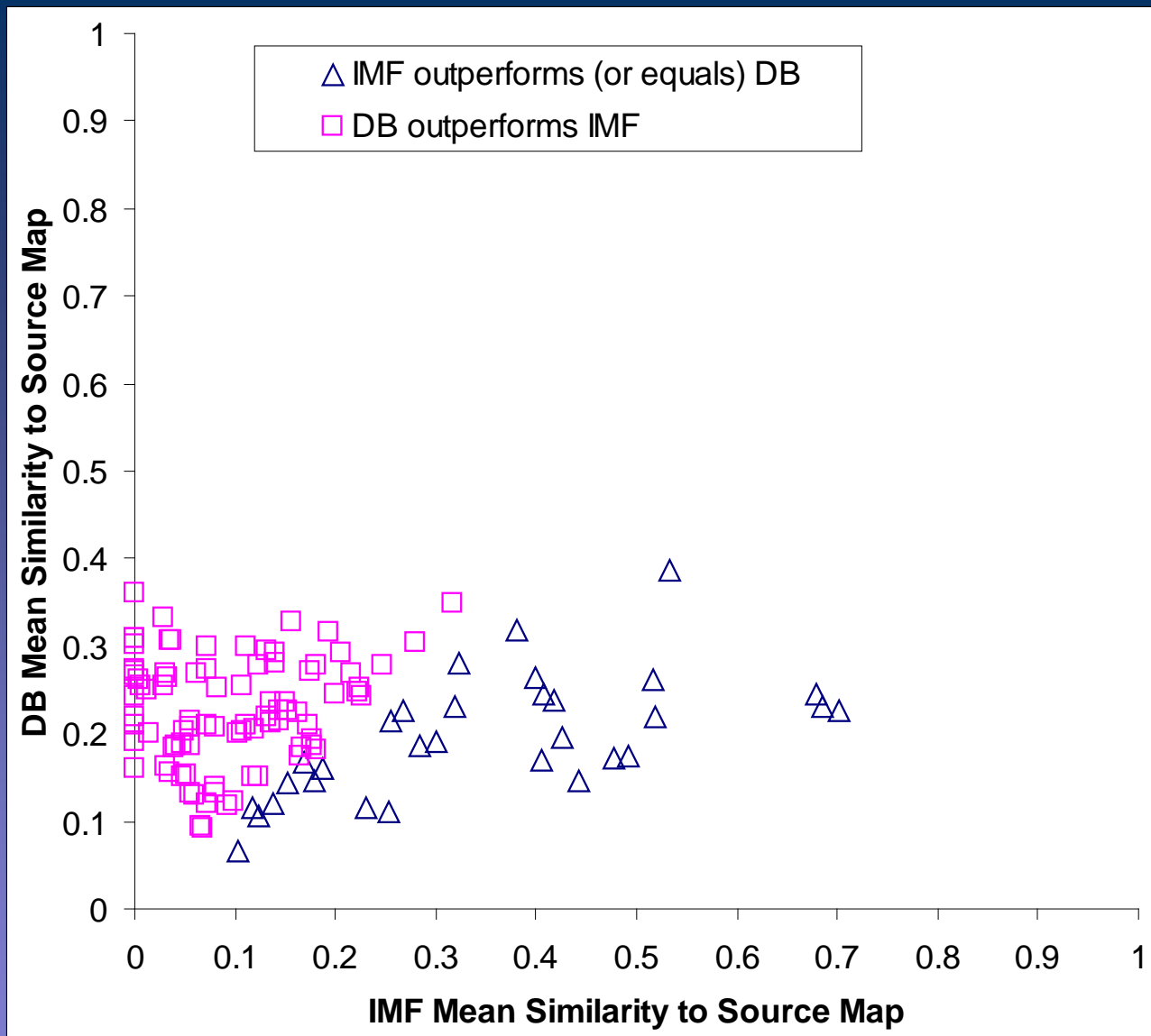
$$S(Q, M, D) = \frac{|(D \cap M) - Q|}{|(D \cup M) - Q|}$$

$Q: t_1 \text{ AND } t_2 \text{ AND } \dots \text{ AND } t_n$

- Compare the performance to a baseline.

DB: use terms with highest Λ value

IMF: use terms with highest "inverse map frequency" value



	N	MEAN	STDEV	SE	95%C.I.
DB	118	0.2196	0.0645	0.0059	(0.2079, 0.2311)
IMF	118	0.1627	0.1563	0.0144	(0.1345, 0.1909)

Conclusions

- We have developed a framework for the dynamic extraction of topic descriptors and discriminators to aid information search in context.
- EXTENDER proactively supports the user's knowledge extension process by suggesting novel but related topics, using an iterative search process.
- We have presented a "bottom up" evaluation of EXTENDER's descriptor and discriminator extraction.
- We have also evaluated EXTENDER's ability to achieve our general desiderata for topic suggestions (coverage, novelty, global coherence) with encouraging results.



Dynamic Extraction of Topic Descriptors and Discriminators: Towards Automatic Context-Based Topic Search

Ana Maguitman, David Leake,

Thomas Reichherzer and Filippo Menczer

Indiana University



Joint project with IHMC

Partially supported by NASA under award No NCC 2-1216