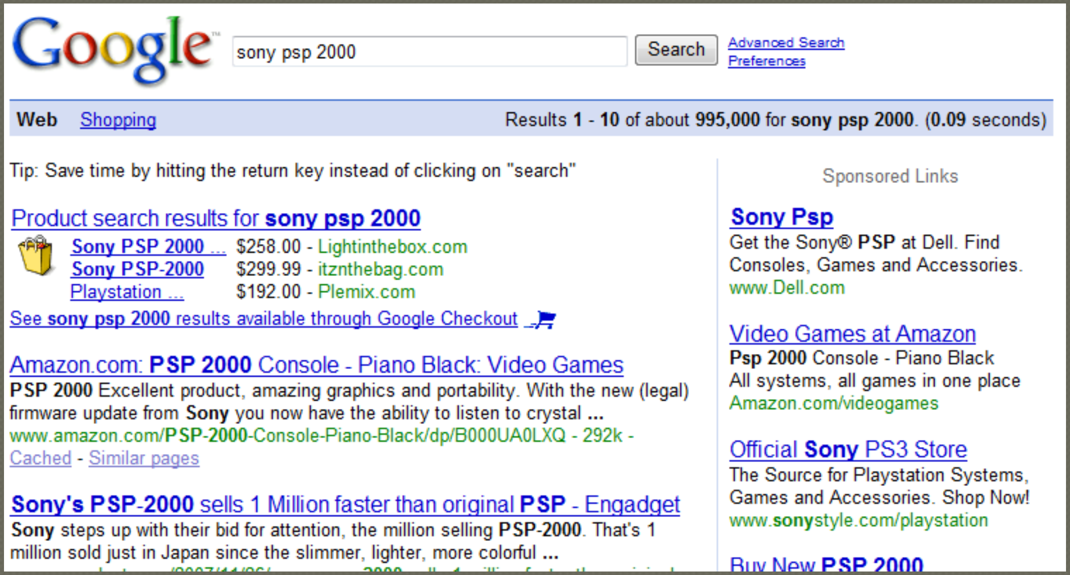
Advertising Keyword Suggestion Based on Concept Hierarchy

Yifan Chen, Guirong Xue and Yong Yu
Apex Data & Knowledge Management LabShanghai Jiao Tong University

Presented by Qiang Yang, Hong Kong Univ. of Science and Technology

Background

- In a Search Engine Company
  - Advertisers bid on keywords
  - Search engine users enter queries
    - Ads w/ keywords that match the query are displayed



Results **1** - **10** of about **995,000** for **sony psp 2000**. (**0.09** seconds)

Tip: Save time by hitting the return key instead of clicking on "search"

Product search results for **sony psp 2000**

Sony PSP 2000 ...	$258.00 - Lightinthebox.com
Sony PSP-2000	$299.99 - itznthebag.com
Playstation ...	$192.00 - Plemix.com

See **sony psp 2000** results available through Google Checkout

Amazon.com: **PSP 2000** Console - Piano Black: Video Games
PSP 2000 Excellent product, amazing graphics and portability. With the new (legal) firmware update from **Sony** you now have the ability to listen to crystal **...**
www.amazon.com/**PSP-2000**-Console-Piano-Black/dp/B000UA0LXQ - 292k -
Cached - Similar pages

Sony's PSP-2000 sells 1 Million faster than original **PSP** - Engadget
Sony steps up with their bid for attention, the million selling **PSP-2000**. That's 1 million sold just in Japan since the slimmer, lighter, more colorful **...**

Sponsored Links

Sony Psp
Get the Sony® **PSP** at Dell. Find Consoles, Games and Accessories.
www.Dell.com

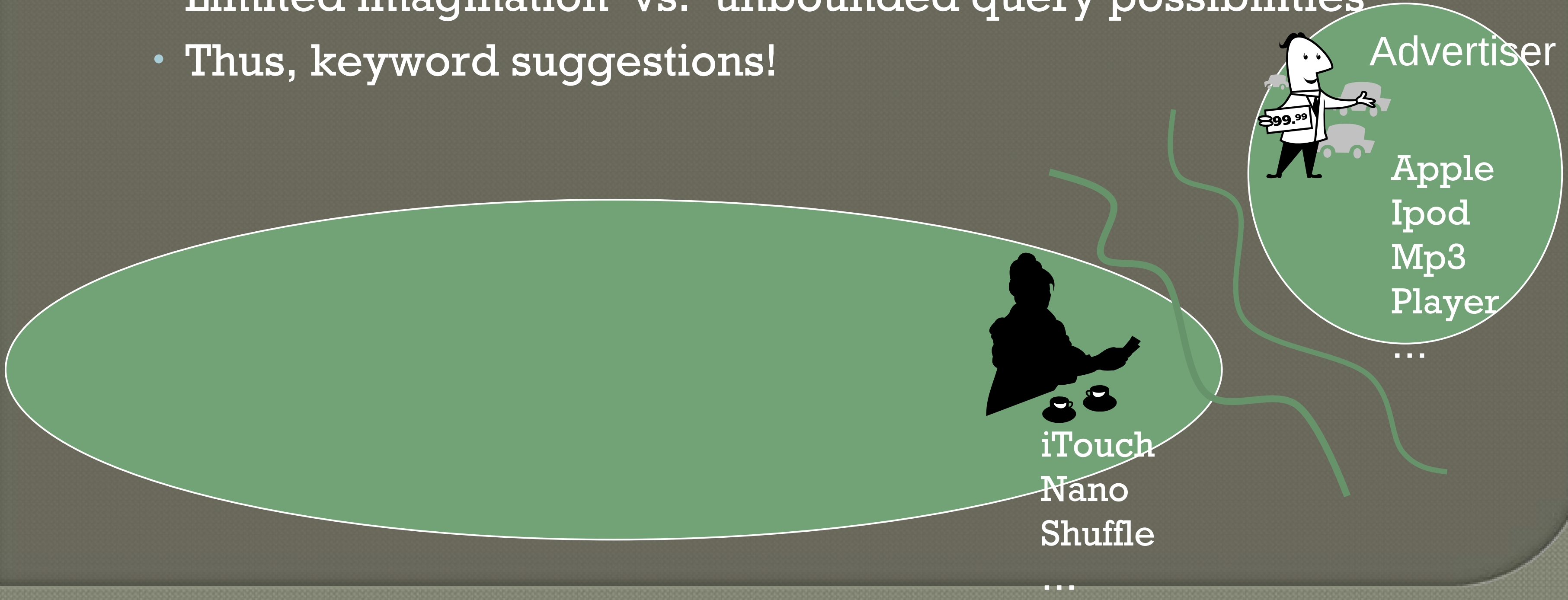
Video Games at Amazon
Psp 2000 Console - Piano Black
All systems, all games in one place
Amazon.com/videogames

Official **Sony** PS3 Store
The Source for Playstation Systems, Games and Accessories. Shop Now!
www.**sony**style.com/playstation

Buy New **PSP 2000**

Background Cont'd

- However…
  - There is a gap between the advertisers and customers,
    - Between Advertisers' vocabulary and customers' vocabulary
    - Limited imagination vs. unbounded query possibilities
    - Thus, keyword suggestions!



Advertiser

Apple
Ipod
Mp3
Player
…

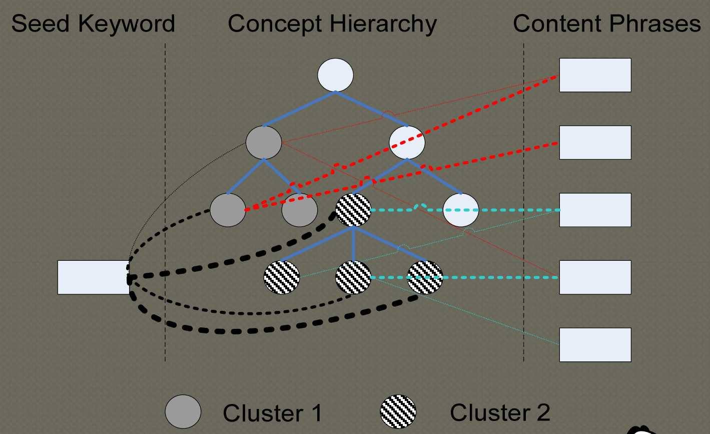
iTouch
Nano
Shuffle
…

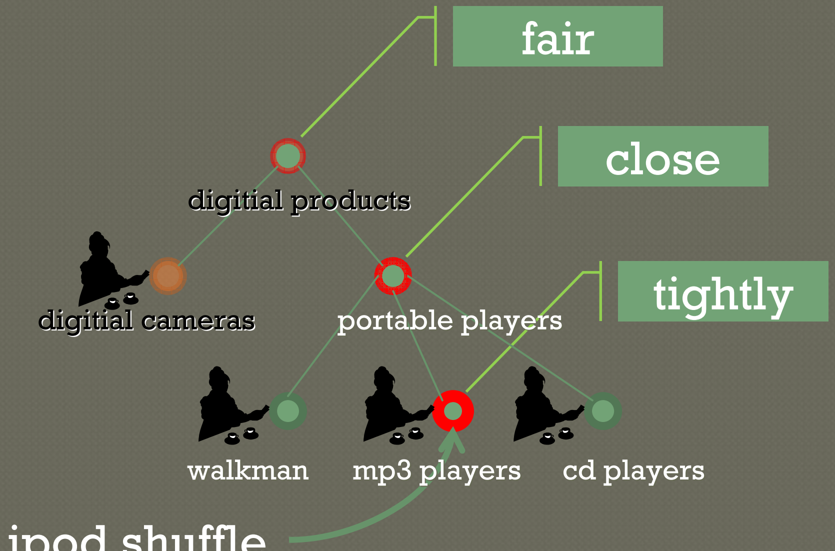
Related work

- Based on query log and advertiser’s logs
  - Adwords from Google, e.g.
  - Find keywords used concurrently
- Find relevant keywords co-occurrence in meta tags
- Based on search-engine result
  - Find near-by phrases in search results

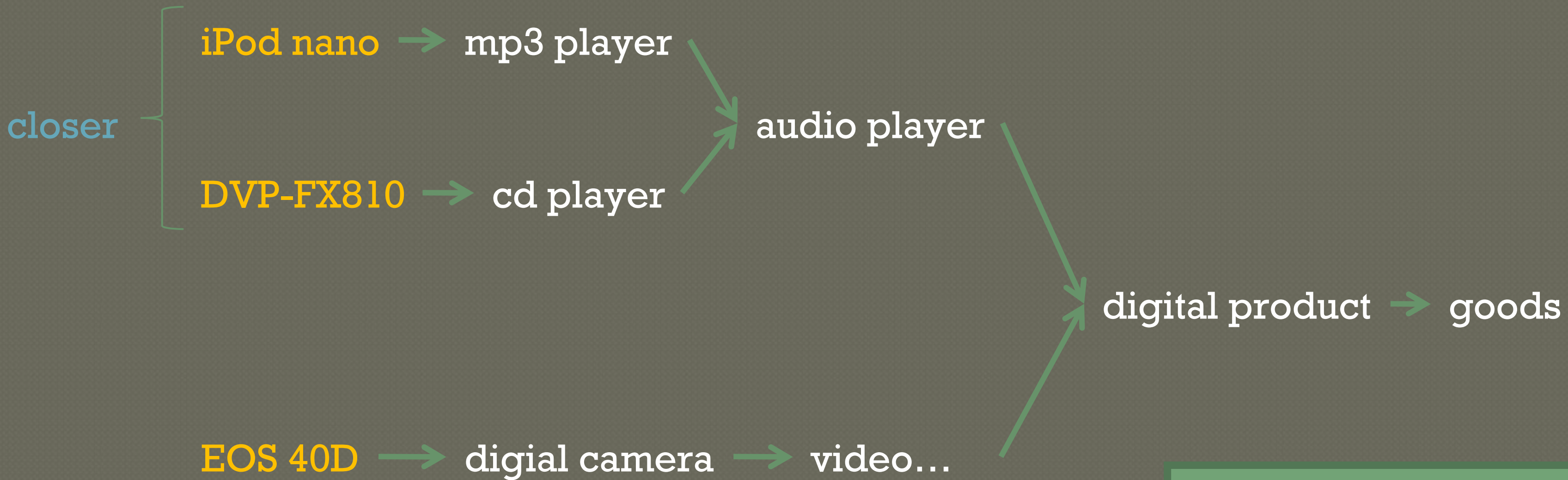
Our approach

- Based on a concept hierarchy
  - Mining the semantic relationships
  - Concentrating on users' real interest





Concept taxonomy induces a distance



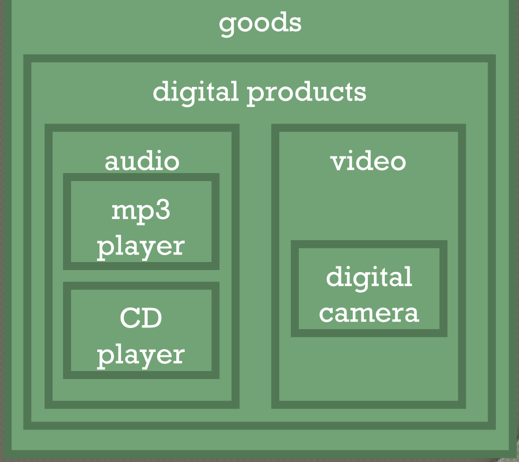
closer

iPod nano → mp3 player → audio player

DVP-FX810 → cd player → audio player

audio player → digital product → goods

EOS 40D → digial camera → video… → digital product



goods

digital products

audio

mp3 player

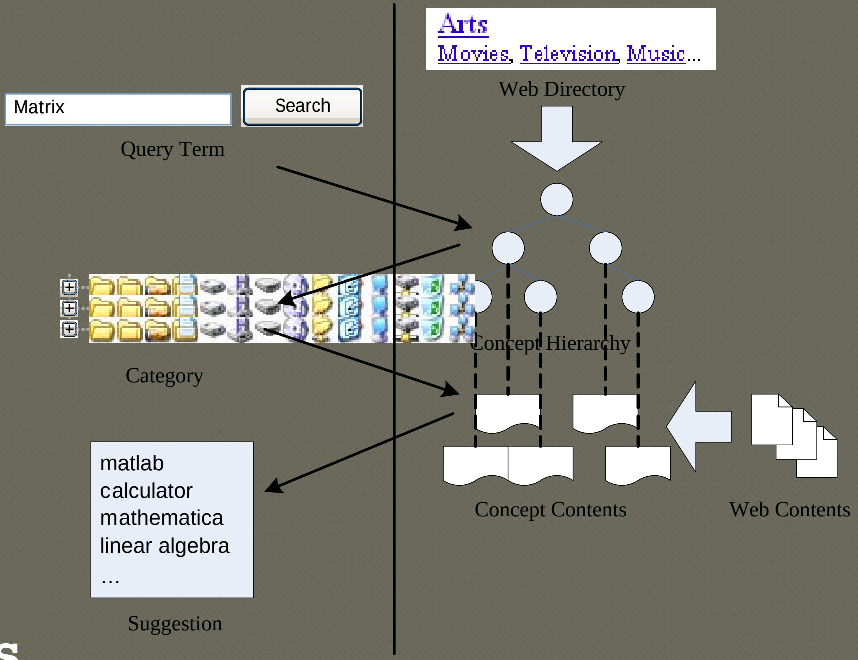
CD player

video

digital camera

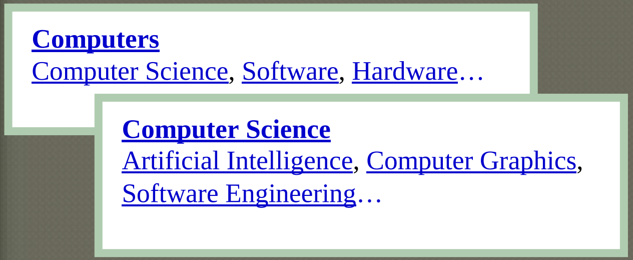
Framework

- Offline:
  - Deriving a Concept Hierarchy
  - Generate keywords
- Online
  - Mapping keywords to generate concept candidates
  - Generate new keywords
  - Categorizing new keywords to concept clusters



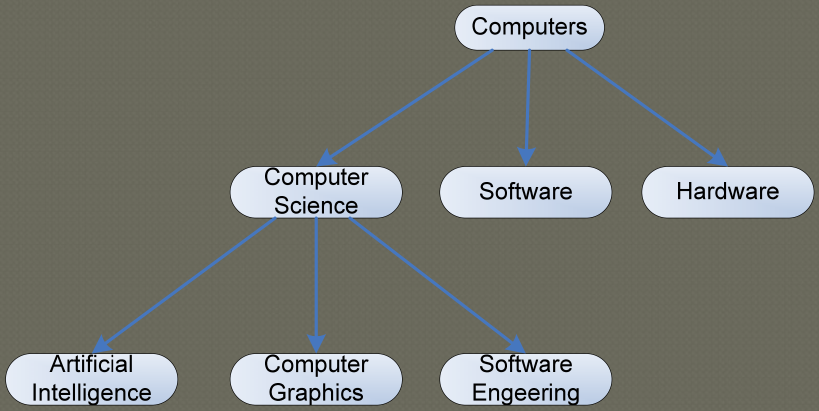
Step 1 – deriving a concept hierarchy

- Built from Web Directories (e.g., ODP)
  - High coverage and accuracy
- Categories as concepts
- Structure as relationship



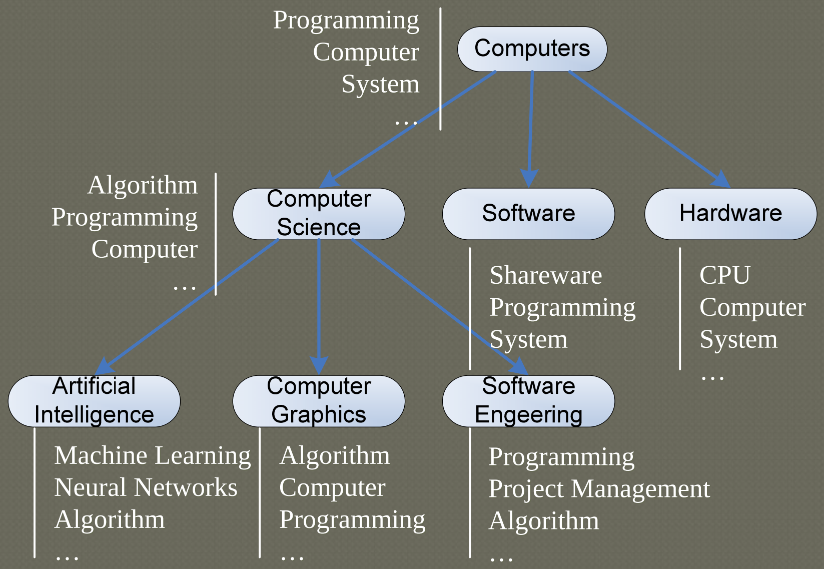
Computers
Computer Science, Software, Hardware…

Computer Science
Artificial Intelligence, Computer Graphics, Software Engineering…



Step 1 – deriving a concept hierarchy

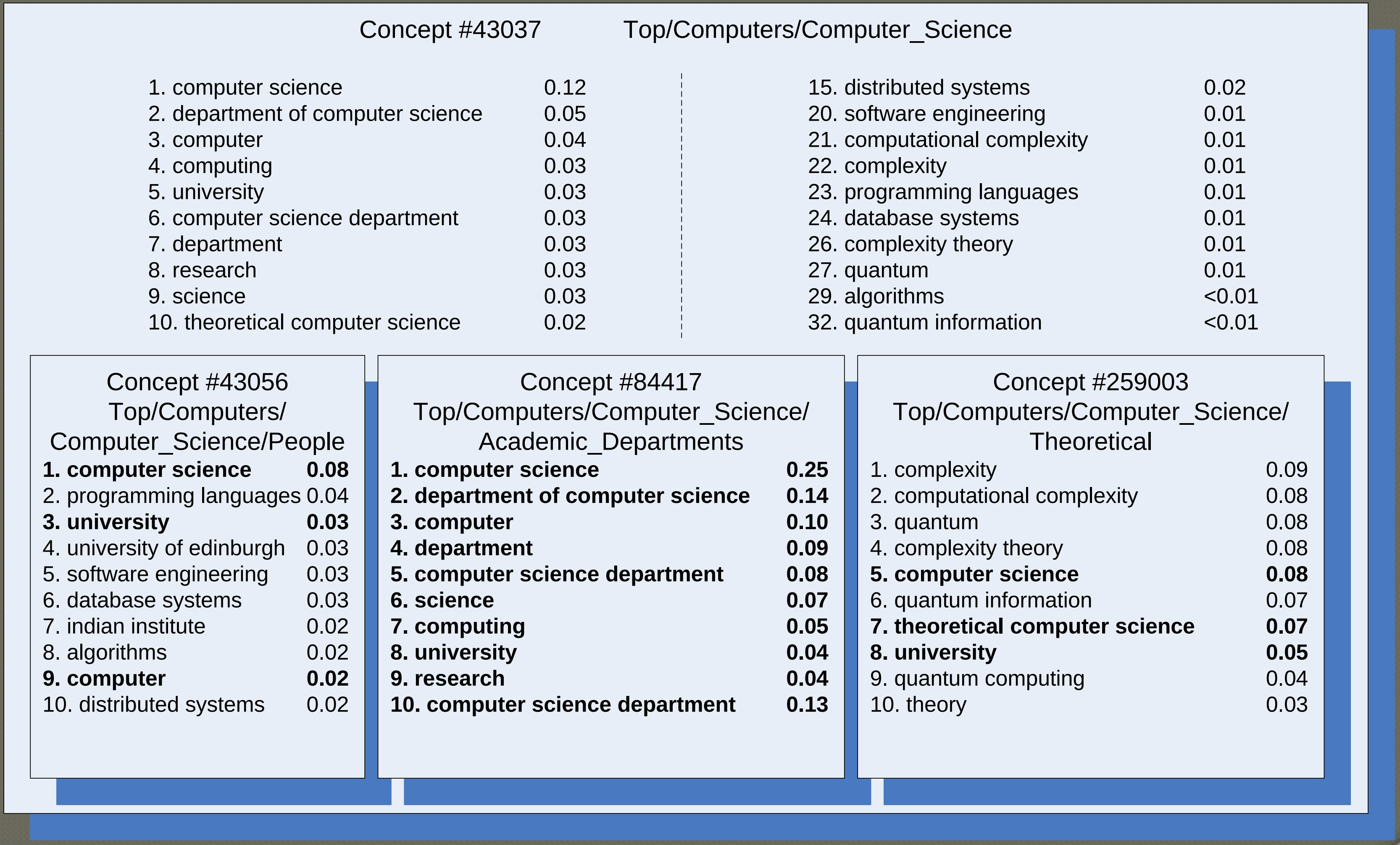
- The meaning of concepts: keywords
  - Phrases gathered from Web Pages under each concept node in taxonomy
    - Keyword extraction
    - Accumulate meaning from sub-concepts



Step 1 – deriving a concept hierarchy

- What keywords are representative for a concept?
- A keyword is good if
  - It is commonly used within the concept (high document frequency)
  - It is seldom used by other concepts (low concept frequency)
  - Similar to TF-IDF, we derive a new keyword evaluation criterion:
    - The Document-Frequency, Inverse-Concept Frequency (DF·ICF) factor

Step 1 – deriving a concept hierarchy



Concept #43037 Top/Computers/Computer_Science

1. computer science	0.12	15. distributed systems	0.02
2. department of computer science	0.05	20. software engineering	0.01
3. computer	0.04	21. computational complexity	0.01
4. computing	0.03	22. complexity	0.01
5. university	0.03	23. programming languages	0.01
6. computer science department	0.03	24. database systems	0.01
7. department	0.03	26. complexity theory	0.01
8. research	0.03	27. quantum	0.01
9. science	0.03	29. algorithms	<0.01
10. theoretical computer science	0.02	32. quantum information	<0.01

Concept #43056
Top/Computers/
Computer_Science/People

1. computer science	**0.08**
2. programming languages	0.04
3. university	**0.03**
4. university of edinburgh	0.03
5. software engineering	0.03
6. database systems	0.03
7. indian institute	0.02
8. algorithms	0.02
9. computer	**0.02**
10. distributed systems	0.02

Concept #84417
Top/Computers/Computer_Science/
Academic_Departments

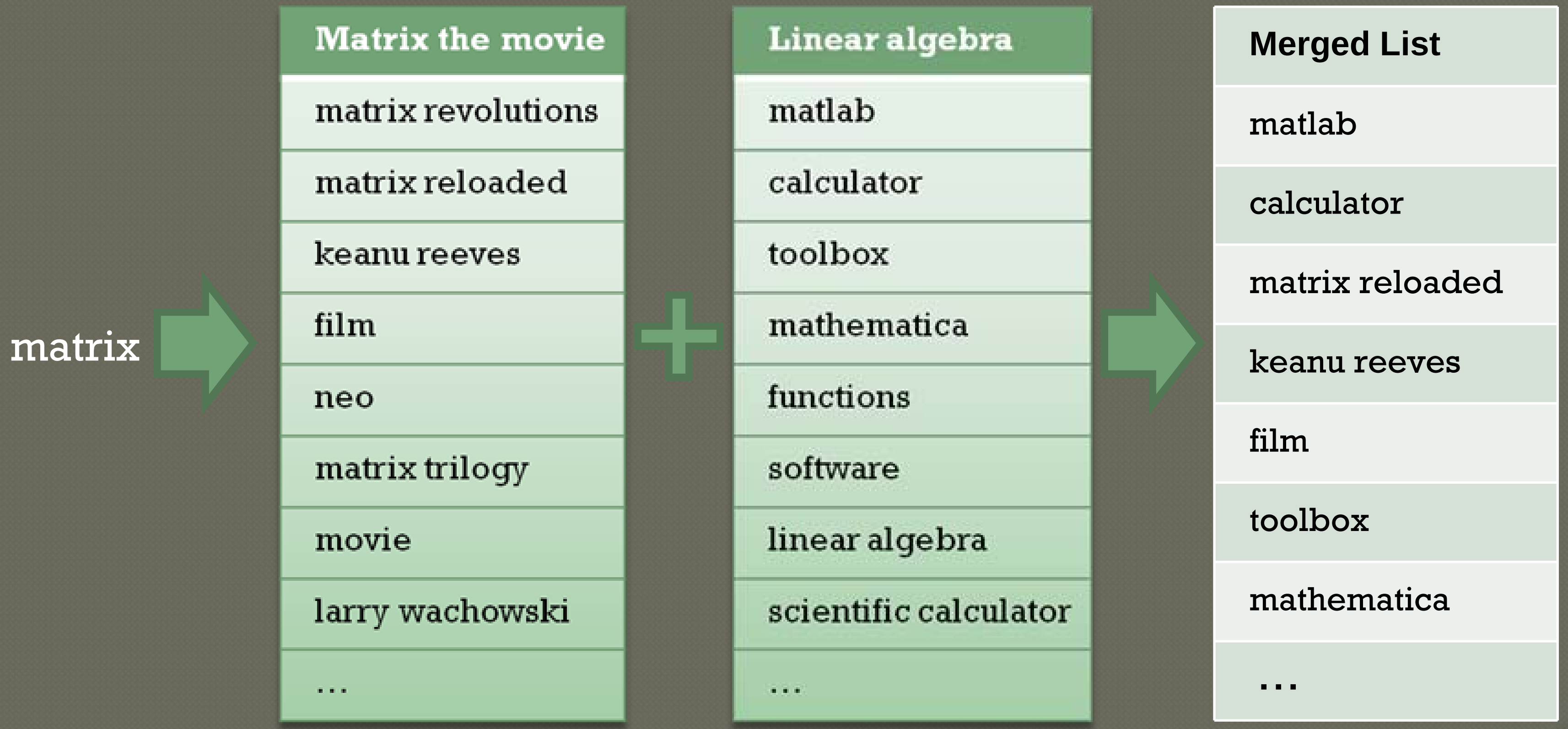
1. computer science	**0.25**
2. department of computer science	**0.14**
3. computer	**0.10**
4. department	**0.09**
5. computer science department	**0.08**
6. science	**0.07**
7. computing	**0.05**
8. university	**0.04**
9. research	**0.04**
10. computer science department	**0.13**

Concept #259003
Top/Computers/Computer_Science/
Theoretical

1. complexity	0.09
2. computational complexity	0.08
3. quantum	0.08
4. complexity theory	0.08
5. computer science	**0.08**
6. quantum information	0.07
7. theoretical computer science	**0.07**
8. university	**0.05**
9. quantum computing	0.04
10. theory	0.03

Keyword Suggestion by merging

- Weighted union: from query q to term t
  - $Sim(q,t)= \sum_c Weight(q->c)*Weight(c->t)$



matrix

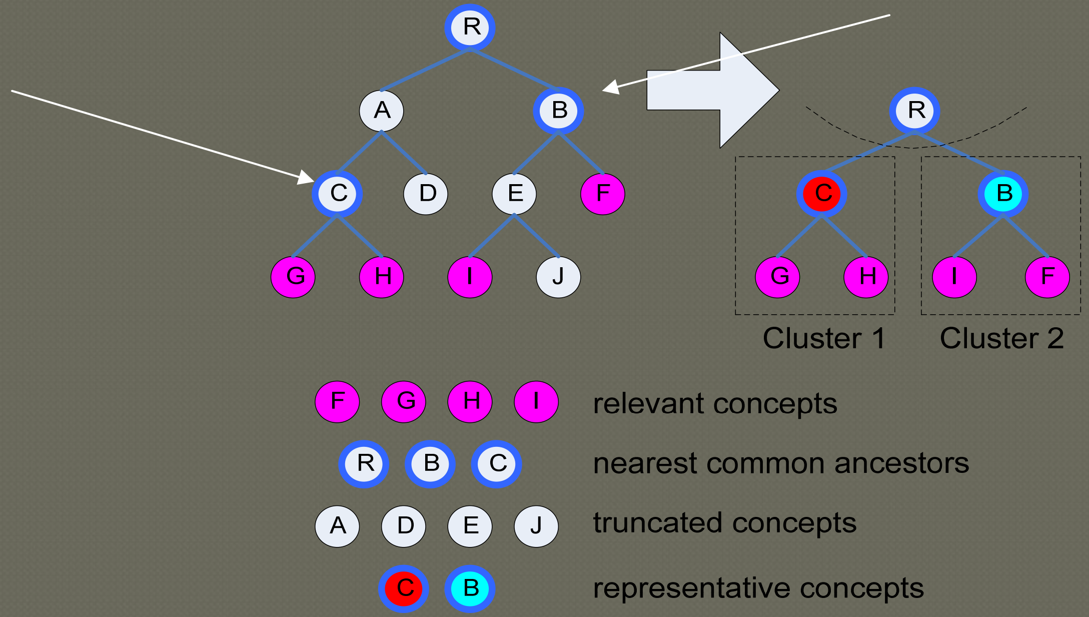
Matrix the movie
matrix revolutions
matrix reloaded
keanu reeves
film
neo
matrix trilogy
movie
larry wachowski
…

Linear algebra
matlab
calculator
toolbox
mathematica
functions
software
linear algebra
scientific calculator
…

Merged List
matlab
calculator
matrix reloaded
keanu reeves
film
toolbox
mathematica
…

Categorizing the Keyword List

- Keyword list Categorization
  - Partition keyword list using concept hierarchy
  - For avoiding concepts from interfering with each other
- Present the Advertisers with categories



Cluster 1

Cluster 2

relevant concepts

nearest common ancestors

truncated concepts

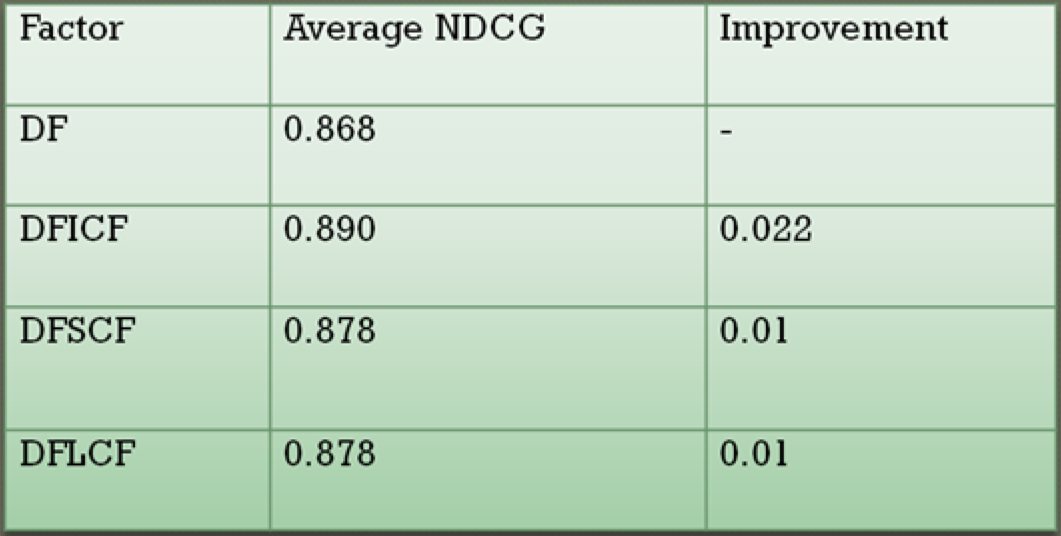
representative concepts

Evaluation

- Dataset
  - 1,306,586 web pages from the 150,446 ODP categories
- Experiments
  - Random Selection of test concepts and keywords
  - 3 labelers are asked to critique the relevance of keywords

Evaluation

- Accuracy of ranked keyword suggestion:
  - Baseline: document frequency (DF)



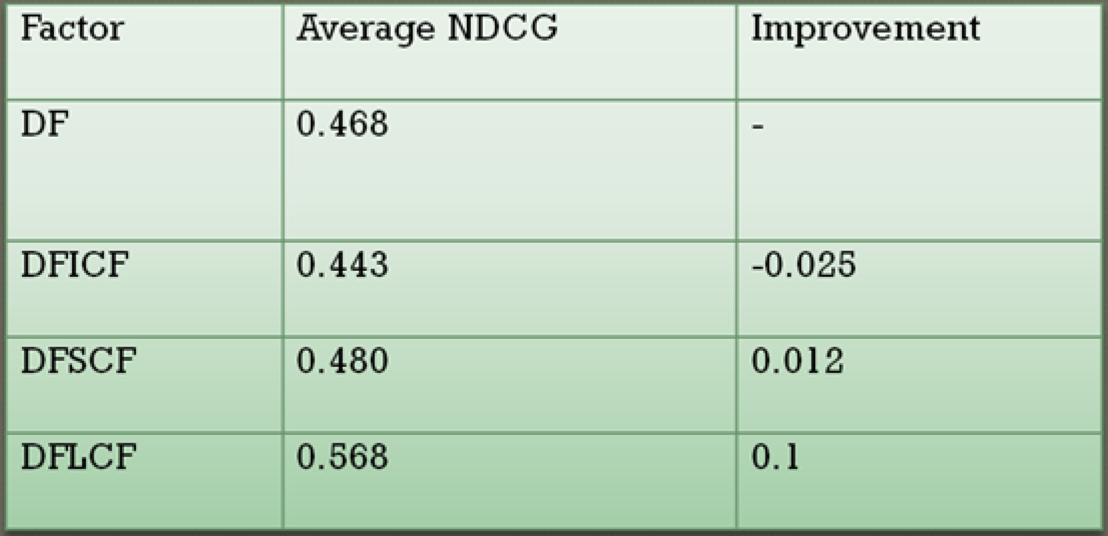
Factor	Average NDCG	Improvement
DF	0.868	-
DFICF	0.890	0.022
DFSCF	0.878	0.01
DFLCF	0.878	0.01

SCF: Sub-category Freq
LCF: Local Concept Frequency

Evaluation

- Accuracy of keyword extraction:
  - Randomly sampled 100 keywords
  - Baseline: DF

SCF: Sub-category Freq
LCF: Local Concept Frequency



Factor	Average NDCG	Improvement
DF	0.468	-
DFICF	0.443	-0.025
DFSCF	0.480	0.012
DFLCF	0.568	0.1

Evaluation

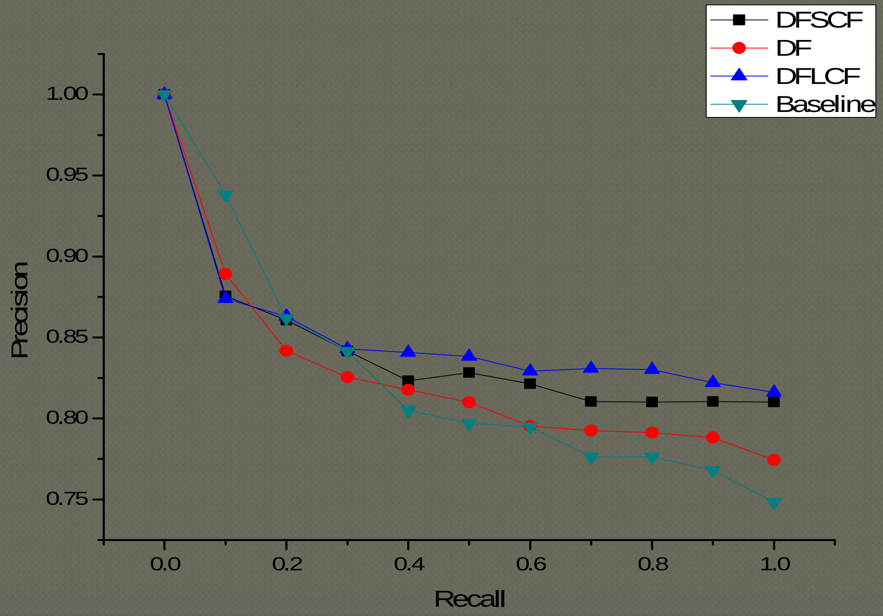
- Completeness of suggestion: all distinct meanings found?
  - Ambiguous queries
  - Baseline: DF

Method	Found	Hit	Relevant	Redundant	Missing
DF	6.4	1.8	0.7	3.9	0.4
DFSCF	7.8	2.4	0.3	5.1	0.2
DFLCF	7.1	2.4	0.3	4.5	0.2

Evaluation

- Disambiguition Performance
  - Suggestion without categorizing
  - Baseline: co-occurrence

Method	Precision	Recall	F1-measure
Baseline	0.74	0.38	0.51
DF	0.77	0.55	0.62
DFSCF	0.81	0.52	0.65
DFLCF	0.82	0.55	0.66



Evaluation cont'd

- Case Study with query "matrix"

Ours		Google	Overture	WordTracker
Top/Arts	Top/Science/Math			
matrix revolutions	matlab	matrix screensaver	the matrix	matrix
matrix reloaded	calculator	matrix reloaded	belief bridging divine matrix miracle space time	the matrix
matrix review	toolbox	matrix revolutions	toyota matrix	matrix reloaded
keanu reeves	mathematica	matrix multiplication	matrix reloaded	toyota matrix
film	functions	matrix inverse	matrix screensaver	matrix soundtrack
neo	software	matrix soundtrack	matrix revolution	matrix revolutions
matrix trilogy	linear algebra	matrix wallpaper	matrix game	matrix theme
movie	scientific calculator	rotation matrix	matrix soundtrack	matrix wallpaper
matrix revolutions review	algebra	matrix code	dot matrix printer	matrix mp3
revolutions	matlab toolbox	matrix revolution	matrix online	matrix screensaver
larry wachowski	biochemistry	matrix inversion	enter the matrix	matrix background
reloaded	graphing	determinant matrix	matrix movie	the matrix reloaded
matrix movies	analysis	trinity matrix	matrix 3d	matrix ping pong
agent smith	calculators	math matrix	enter guide matrix official strategy	matrix code
laurence fishburne	data analysis	symmetric matrix	matrix mris	matrix movie
andy wachowski	computer algebra system	matrix properties	matrix hair product	
morpheus	department	rank matrix	matrix wallpaper	
neo and trinity	linear	algebra matrix	matrix trilogy	
matrix movie	archaeology	matrix product	matrix neo path	
matrix revisited	molecular biology	matrix hacking	matrix shampoo	

Conclusion

- Our work
  - A novel approach for advertising keyword suggestion.
  - Key idea: associate concepts with keywords tightly.
  - Disambiguation to avoid interfering.
- Future work
  - Beyond using web pages
  - keep concept hierarchy up-to-date
    - Automatic content reinforcement from web content such as query logs