Which Concepts Are Worth Extracting?

Arash Termehchy#, Ali Vakilian*, Yodsawalai Chodpathumwan*, Marianne Winslett*

#Oregon State University

*University of Illinois at Urbana-Champaign

The vast majority of data is not structured.

Scientific articles, HTML pages, …

<article id=1>

Michael Jordan is a former American professional basketball player ...

</article>

<article id=2>

Michael Jordan is a full professor at the University of California, Berkeley ...

</article>

<article id=3>

The Michael Jordan's sculpture is in the front of Union Center …

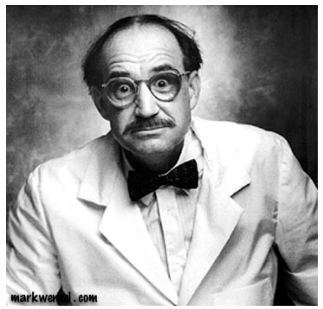
</article>

<article id=4>

All six championship teams of Chicago Bulls were led by Michael Jordan and …

</article>

Users



Keyword query

Michael Jordan Statue

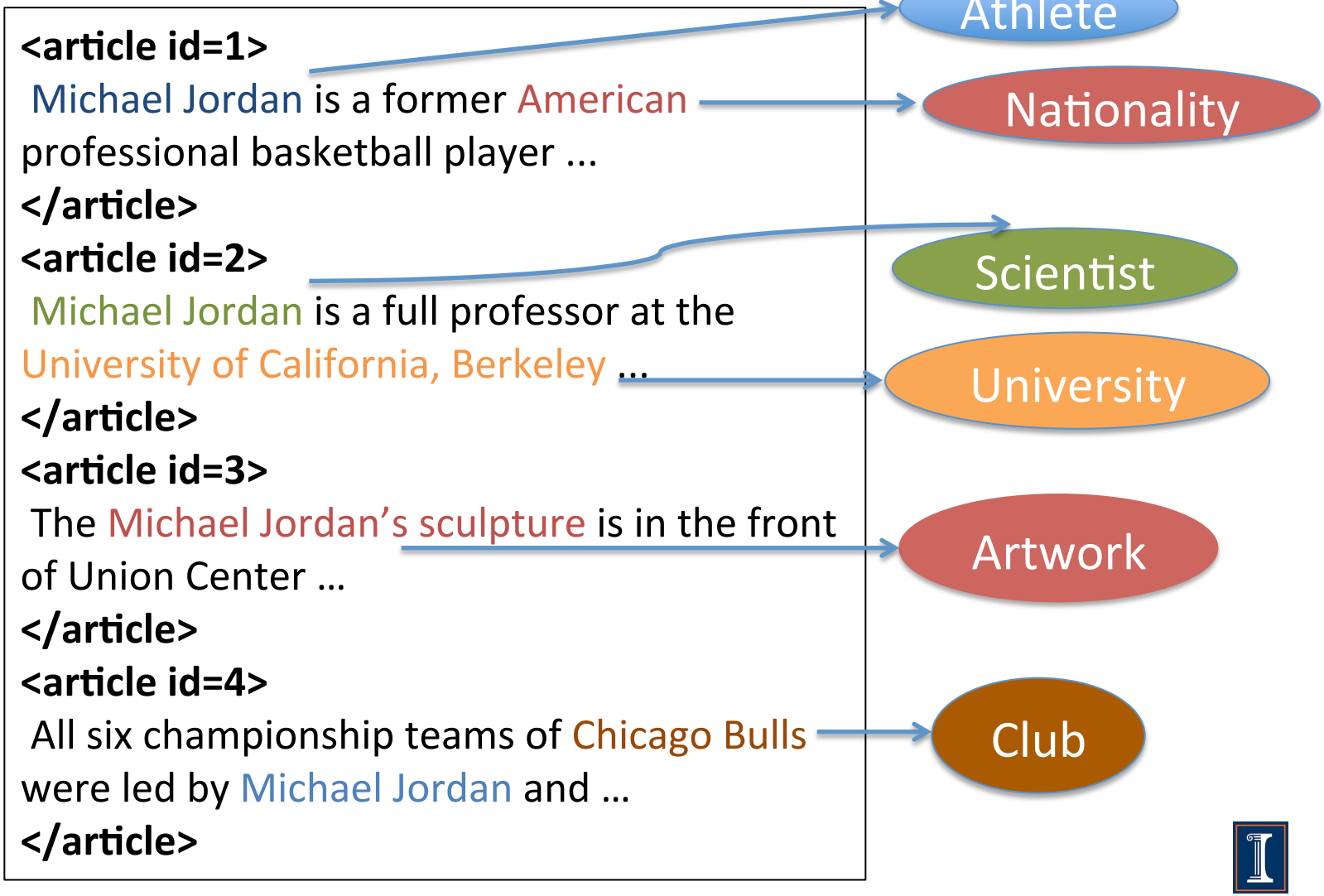
Ranked list

article id=1 ✗
article id=4 ✗
article id=2 ✗
article id=3 ✓

poor ranking quality = frustrated user

Information extraction comes to the rescue!

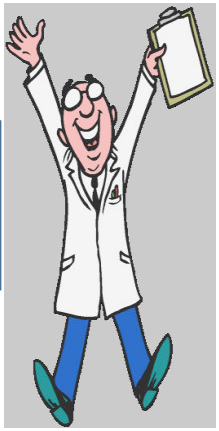
It extracts and organizes the concepts that appear in unstructured information.



Users can submit more structured queries.

<article id=1>
Michael Jordan is a former American professional basketball player ...
</article>
<article id=2>
Michael Jordan is a full professor at the University of California, Berkeley ...
</article>
<article id=3>
The Michael Jordan's sculpture is in the front of Union Center …
</article>
<article id=4>
All six championship teams of Chicago Bulls were led by Michael Jordan and …
</article>

Artwork(*Michael Jordan*)



Ranked list

article id=3 ✓

Artwork

The instances of each concept is extracted by a program called *extractor.*

It is costly to develop, execute, and maintain an extractor.

- Developing thousands of rules; finding, selecting, and extracting relevant features; … . Harder in specific domains like medicine.
- Executing an extractor may take several days.
- Re-writing and re-executing extractors as the underlying data set evolves.

Different concepts have different costs:

Email** versus **Scientist

OSU Oregon State University

Most domains have a large number of concepts.

Plant Ontology (plantontology.org): thousands concepts.

- An enterprise has limited amount of resources.
- Most users cannot wait for a fully extracted data set.

- **We have to extract a subset of concepts in the domain: a *conceptual design* for the data.**

OSU Oregon State University

Cost effective conceptual design problem

Conceptual design S is *cost effective* if

- $\sum_{C \in S} w(C) \leq B$

Cost function

Fixed budget

- *S* improves the ranking quality of answering queries more than other feasible designs.

Currently guided by intuition.

We have to quantify this:
the benefit of a design

OSU Oregon State University

Conceptual design *S* directly helps answering queries whose concepts are in *S*.

<article id=2>
Scientist
Michael Jordan is a full professor at the University of California, Berkeley ...
</article>
<article id=3>
The Michael Jordan statue is a bronze sculpture of the basketball player ...
</article>

Scientist(*Michael Jordan*)

article id=2 ✓

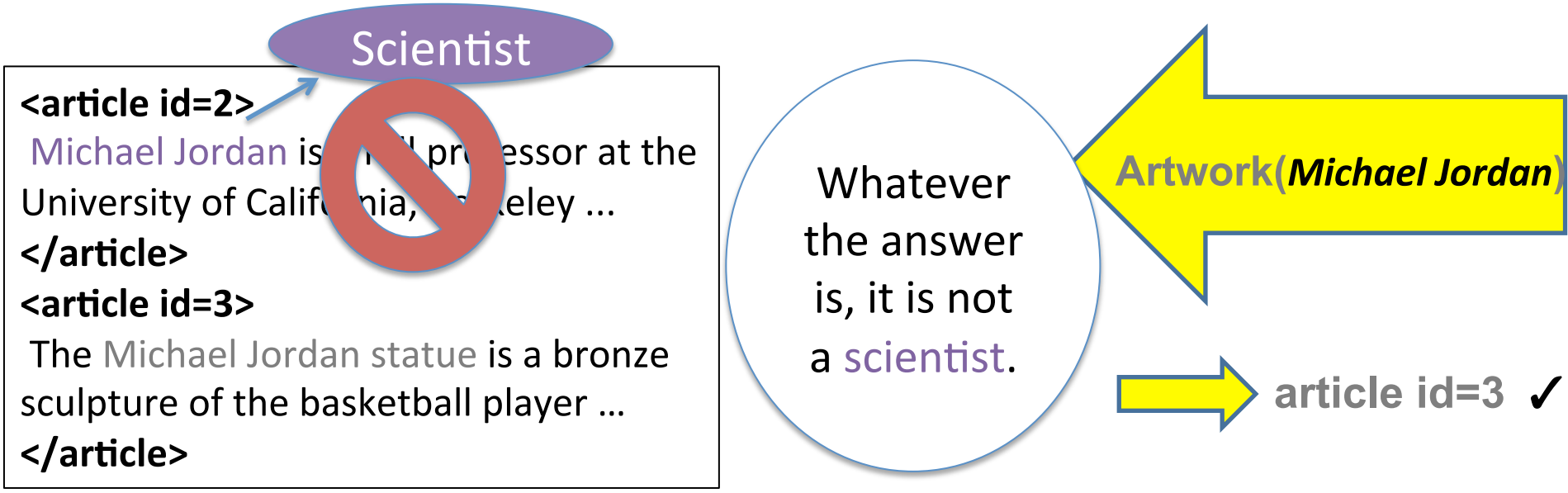
The portion of queries whose concepts are *C*

The accuracy of extracting *C*

$$\sum_{C \in S} u(C)ac(C)$$

What about queries whose concepts are not in the design?

If the concepts are mutually exclusive concepts, we can ignore the instances of the concepts in the design.



Concepts are mutually exclusive.

Generally, the concepts with more instances in the collection are more likely to appear in the top-K answers.

The portion of queries whose concepts are *C*

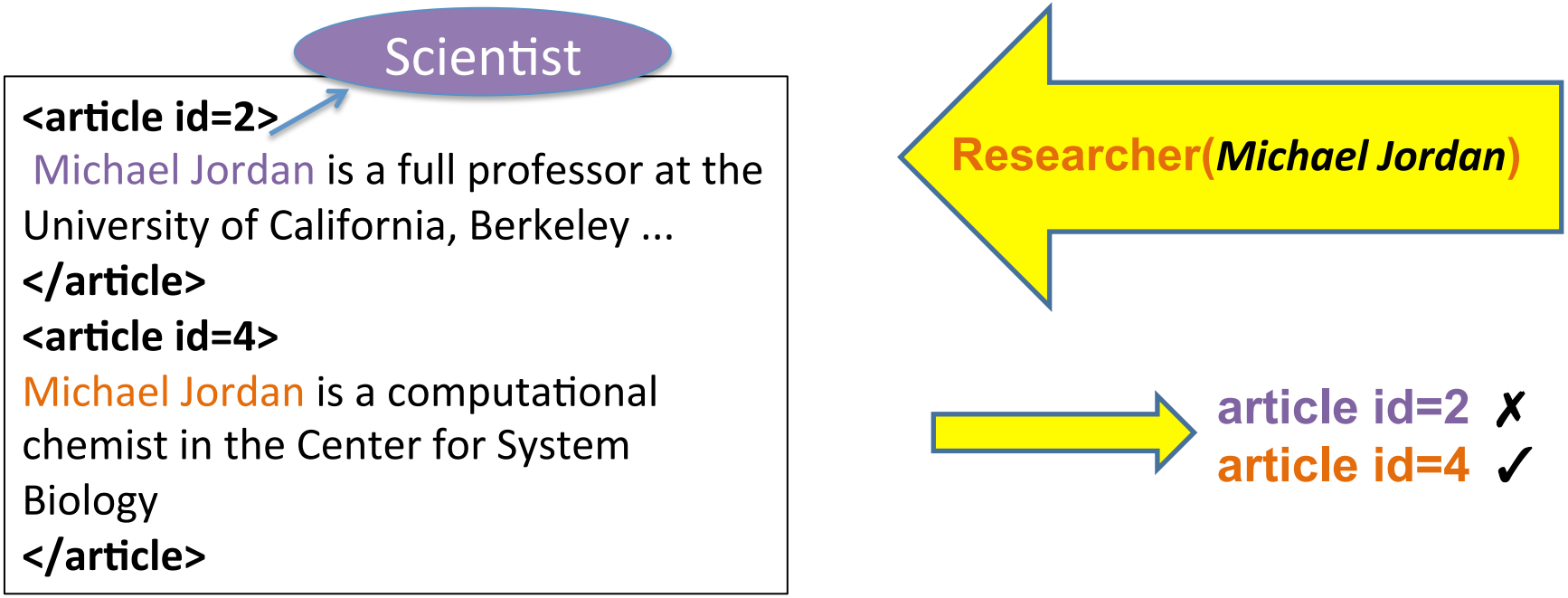
$$\sum_{C \notin S} u(C) \frac{d(C)}{\sum_{E \notin S} d(E)}$$

The portion of entities in the collection that belong to C, *frequency of C*

The portion of entities in the collection whose concepts are not in *S*.

What about queries whose concepts are not in the design?

If there is no constraint regarding the overlap of concepts, we have to consider all concepts in the data.



What about queries whose concepts are not in the design?

If there is not constraint regarding the overlap of concepts.

The portion of queries whose concepts are *C*

$$\sum_{C \notin S} u(C)d(C)$$

The portion of entities in the collection that belong to C, *frequency of C*

Cost effective conceptual design problem

Given a fixed budget *B*, cost function *w*, find conceptual design S such that $\sum_{C \in S} w(C) \leq B$ and

Case 1) If the concepts are mutually exclusive concepts

$$Max \sum_{C \in S} u(C)ac(C) + \sum_{C \notin S} u(C) \frac{d(C)}{\sum_{C \notin S} d(C)}$$

Case 2) No constraints regarding the overlap of concepts:

$$Max \sum_{C \in S} u(C)d(C) + \sum_{C \notin S} u(C)d(C)$$

OSU Oregon State University

The problem is NP-hard in both cases in the number of concepts in the domain.

We propose two efficient approximation algorithms:

APM: prefers concepts that are used more often in queries.

	Approximation ratio
No constraints regarding overlap	1 + ε
Mutually exclusive concepts	2 + ε

AAM: prefers concepts that are used more often in queries and do not have a lot of instances in the collection.

- Approximation ratio of 1+ε over mutually exclusive concepts.

OSU Oregon State University

We evaluate our model and algorithms over Wikipedia, four sets of concepts from YAGO with 7 – 87 concepts, and 1737 queries from MSN.

How well benefit maximization finds the designs with maximum ranking qualities?

Budget (0-1)	Oracle	Benefit Maximization
0.1	0.190 / 0.442	**0.190 / 0.442**
0.2	0.208 / 0.513	**0.208 / 0.513**

precision@3 and MRR (more results in the paper)

OSU Oregon State University

Ranking quality of the designs delivered by our algorithms.

Budget (0 – 1)	**M2** (mutually exclusive concepts)	
	APM	**AAM**
0.1	0.221 /0.517	**0.240 /0.641**
0.2	0.223 /0.532	**0.240 /0.643**

precision@3 and MRR (more results in the paper)

AAM approximates the optimal design more effectively than APM for mutually exclusive domains.

OSU Oregon State University

Running times of approximation algorithms (in minutes)

Algorithm	**M2** (76 concepts)
APM(ε=0.001)	**12**
AAM(ε=0.3)	**5**

precision@3 and MRR (more results in the paper)

- Reasonable for a design time process.

- We can decrease their running times by picking smaller values for ε without considerably affecting their effectiveness.

OSU Oregon State University

Related Research Problems

- *Classic database conceptual design*
  - It does not consider the issue of cost effectiveness.
- *Optimizing information extraction programs*
  - The design is fixed. It deals with issues raised during running time.
  - It optimizes mainly the execution time.

OSU Oregon State University

Conclusion and future work

- Since extracting concepts are costly, we should select a cost effective design for our data.
- We formalized how a conceptual design improves the ranking qualities of answering queries.
- We provided efficient and effective algorithms to select a conceptual design for a collection.
- We plan to solve the problem for other types of relationships between concepts such as IS-A.

OSU Oregon State University