Mining the Web for Relations

Presented by Hongbin Lu

Paper Overview

- This paper proposed a method to discover relations on the Web. Relations means the way different pieces of information are related as they presented on the Web.
- Examples of relations (author, title), (acronym-expansion), ...
- One way is to study patterns of occurrences of related phrases in web documents in order to identify relations between them. We call these the duality problems of the web.
- This paper defined and formalized the duality problem of relations and proposed a general approach to solve those kinds of problems.

Duality Problems

- Duality problems are materialized in trying to identify two sets of inter-related concepts.
- In the WWW, duality exists in two forms:

1. One induced by static link topology.

2. The other occurring, in the text of web document, in the form of relations and patterns.

Problem 1: Authoritative Source in WWW

- Given search query by user, we want to find the most authoritative pages in WWW.

- Assume an index-base search engine is provided for us. That engine search the Web, index Web pages, and build and store huge keyword-based indices that help locate sets of Web pages containing certain keywords.

- If the query is very board, such as "Java", the index-base search engine will return millions of pages.

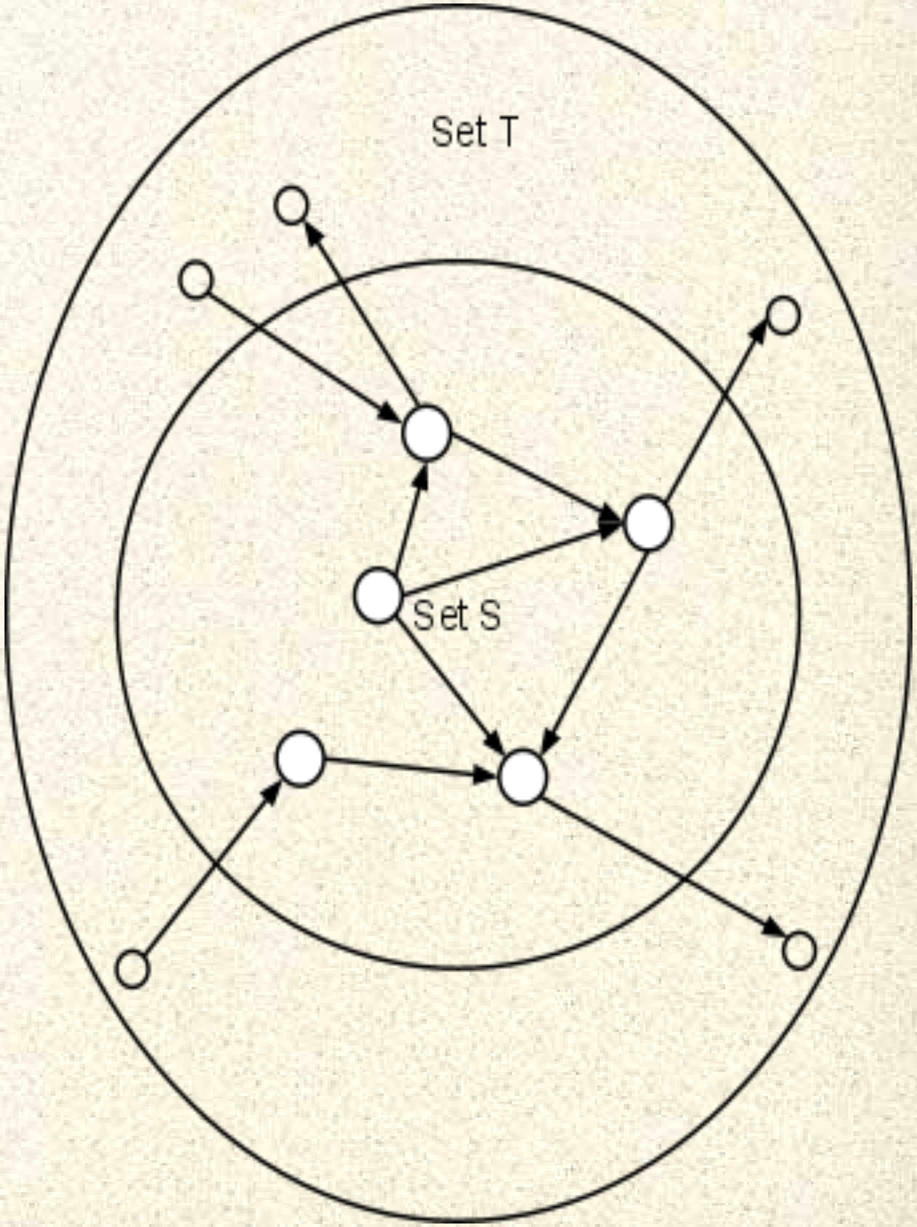
- We want to find the most authoritative page among them.

Reduce the amount of pages

- Pick the top 200 pages return by index-based search engine and let the set of pages as S.

- Expands the set S to a larger set T by adding in any pages that point to, or are pointed to by, any page in S.

- In practice, set S contains almost all the authoritative pages. The set S is called base set.

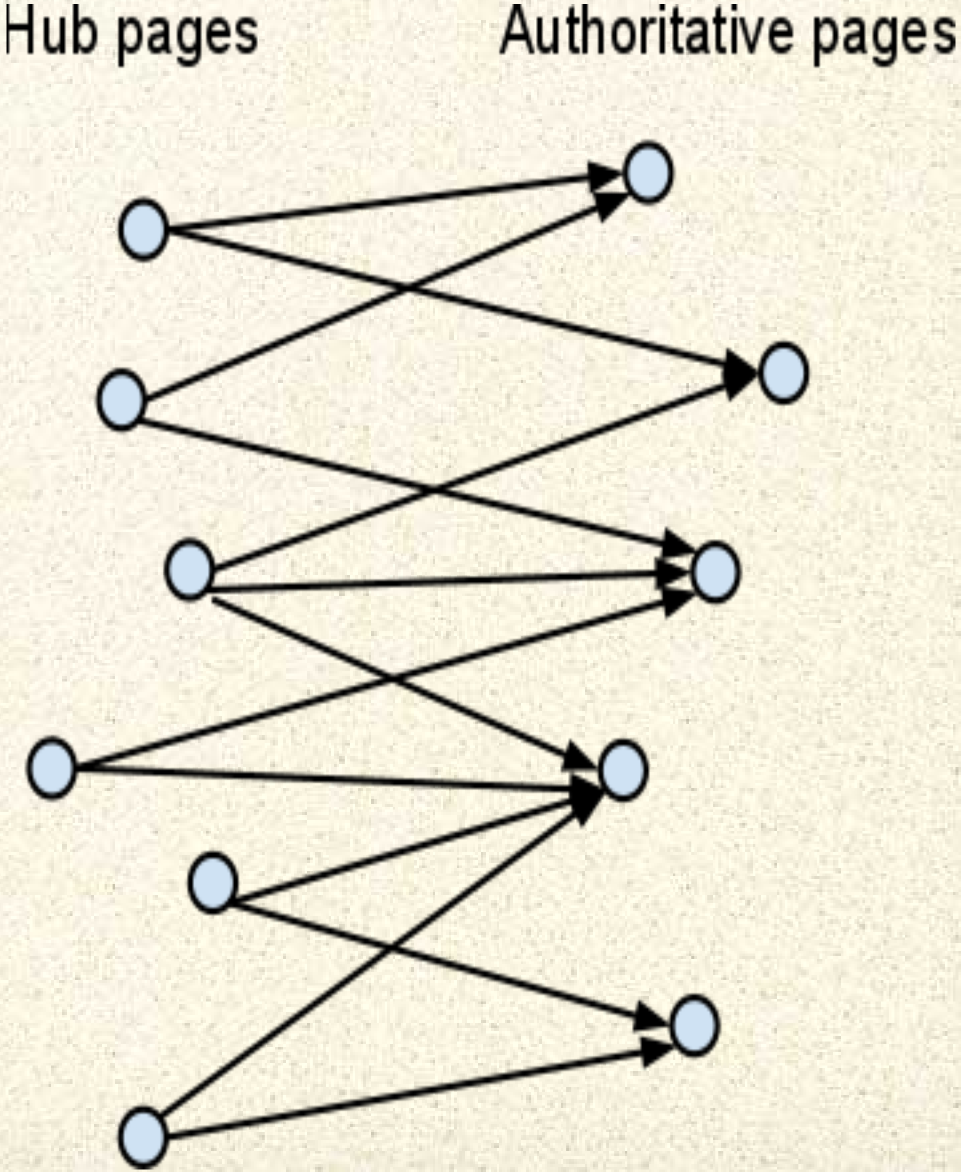


Identify authoritative pages from base set

- In practice, authoritative pages will be pointed by a certain set of pages, which is called "hub".

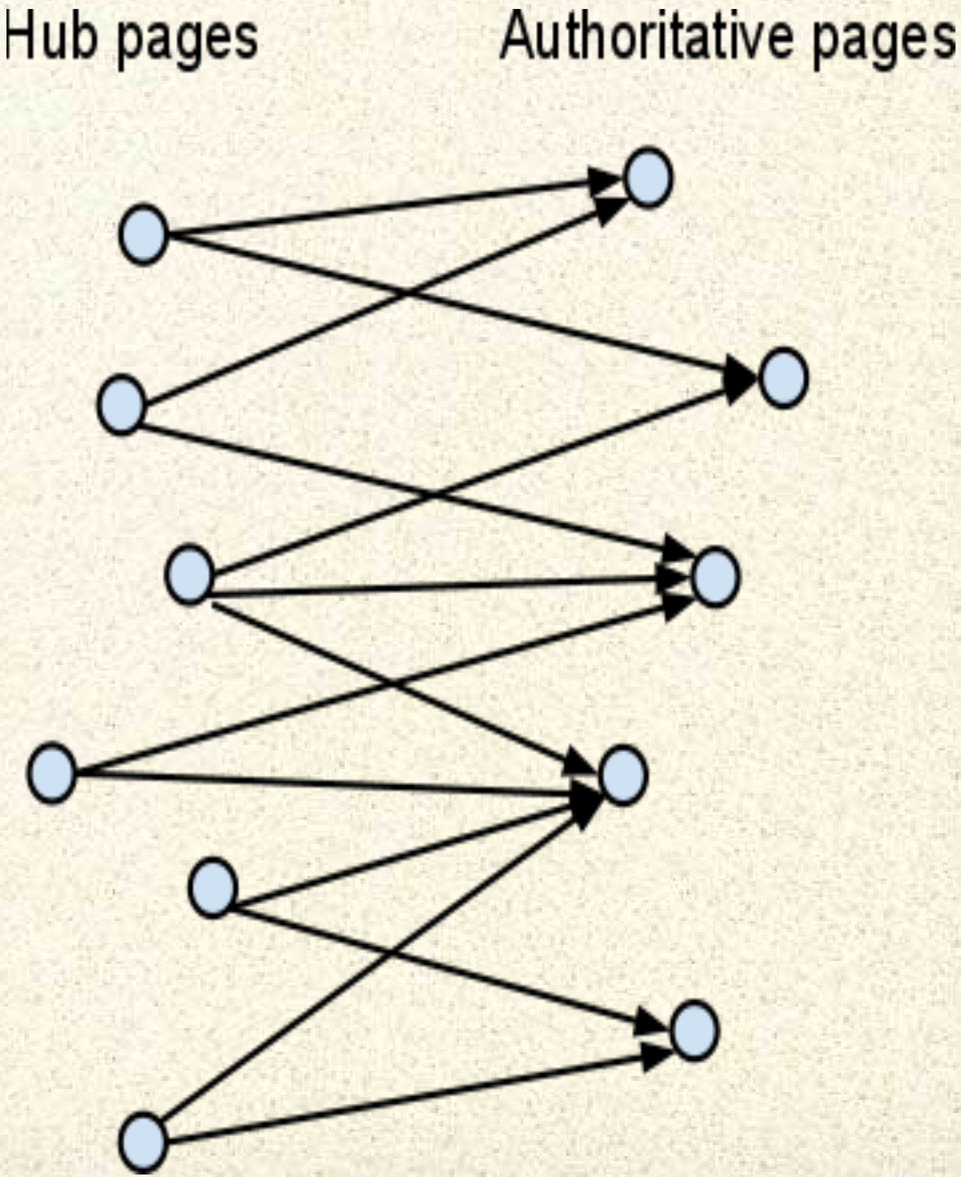
- A good hub page will point to a large amount of authoritative pages.

- Therefore, we can find the set of authoritative pages by finding a good set of hub pages.



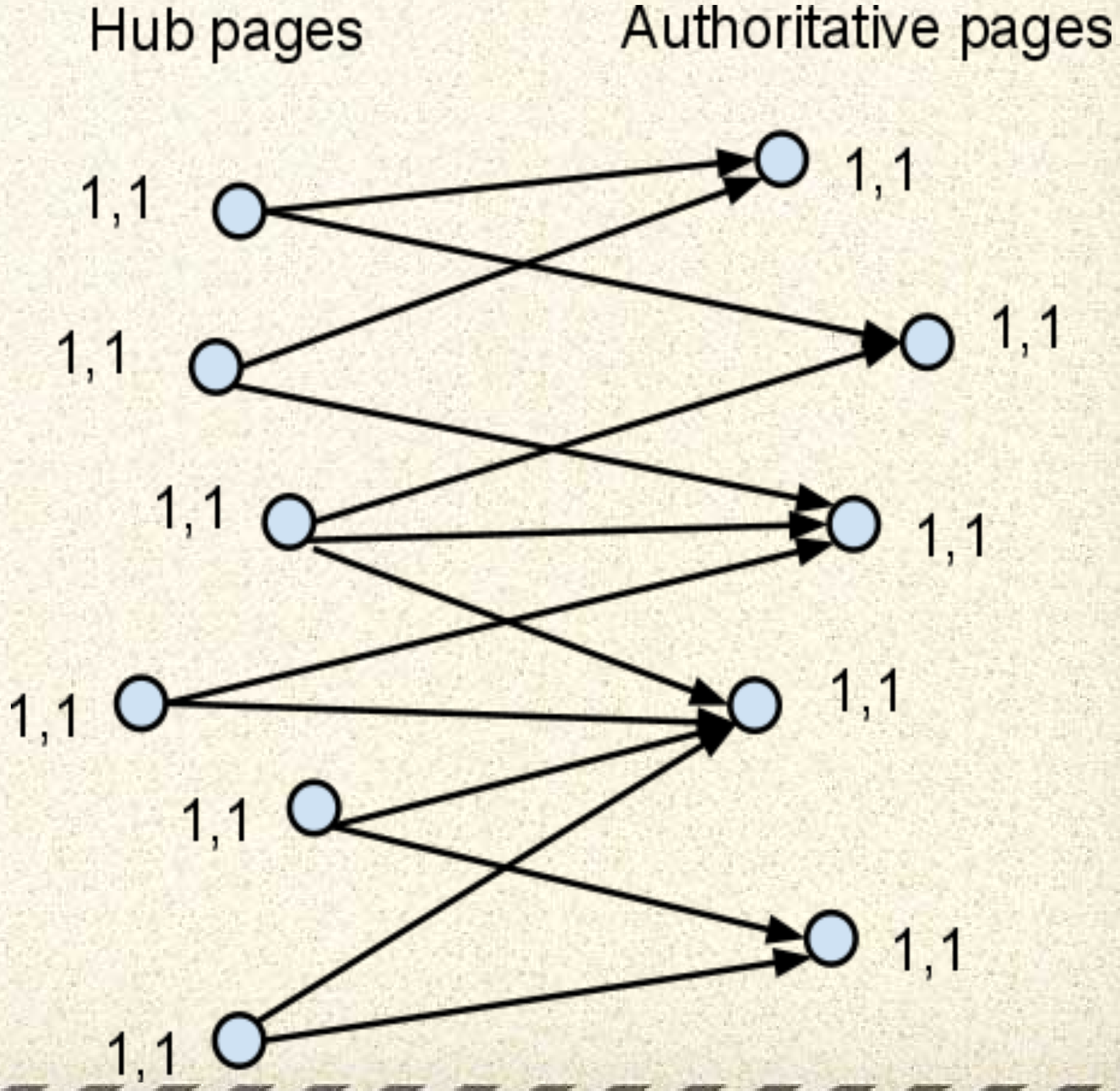
Identify hubs pages from base set

- A Good hub pages is the pages pointing to many authoritative pages.
- The problem is now back to finding a set of authoritative pages.
- Hubs pages and authoritative pages can mutual enforce each other.



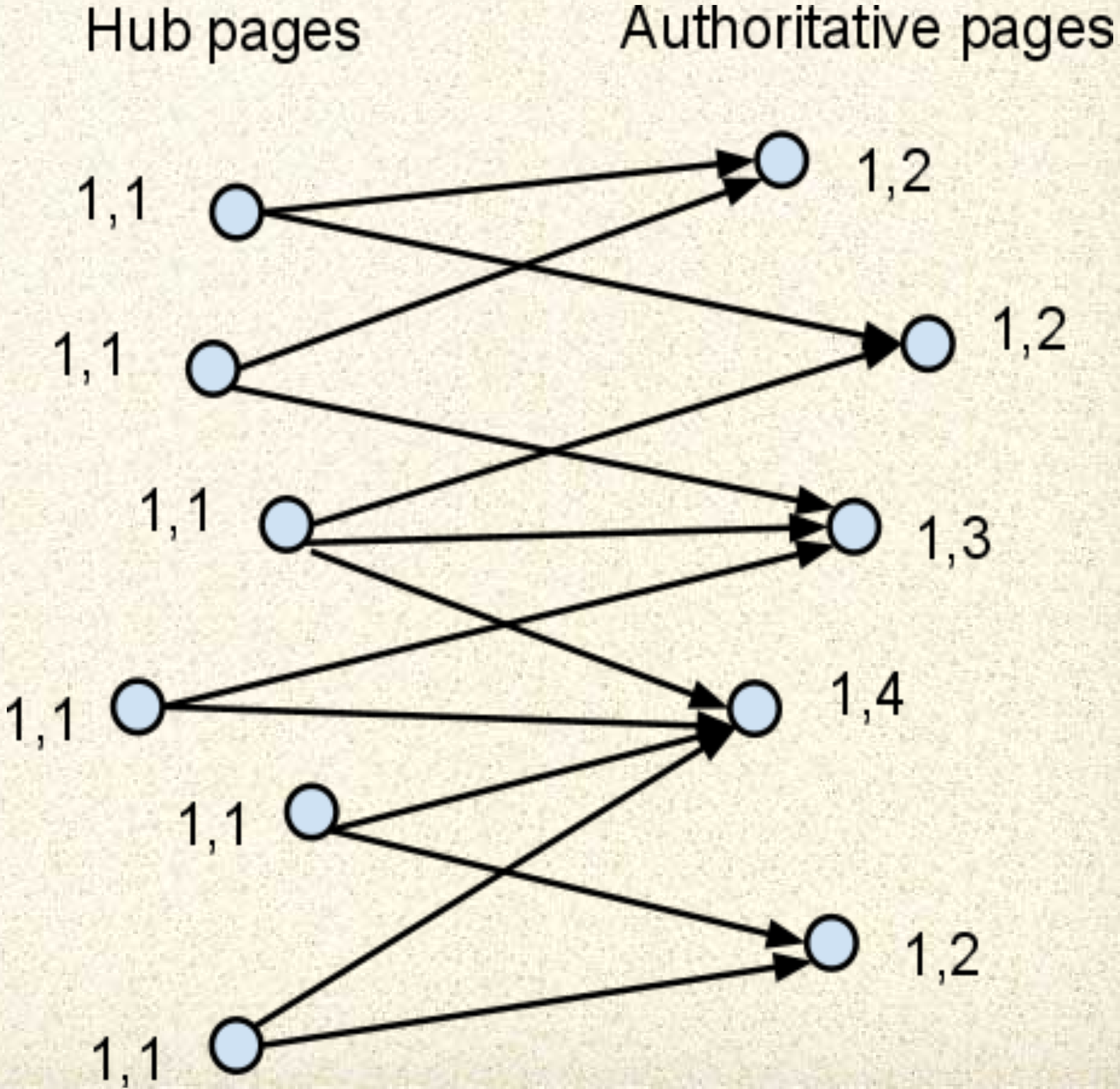
An iterative approach-1

- Associate with each page p a hub weight h(p) and an authority weight a(p), all initialized to 1.



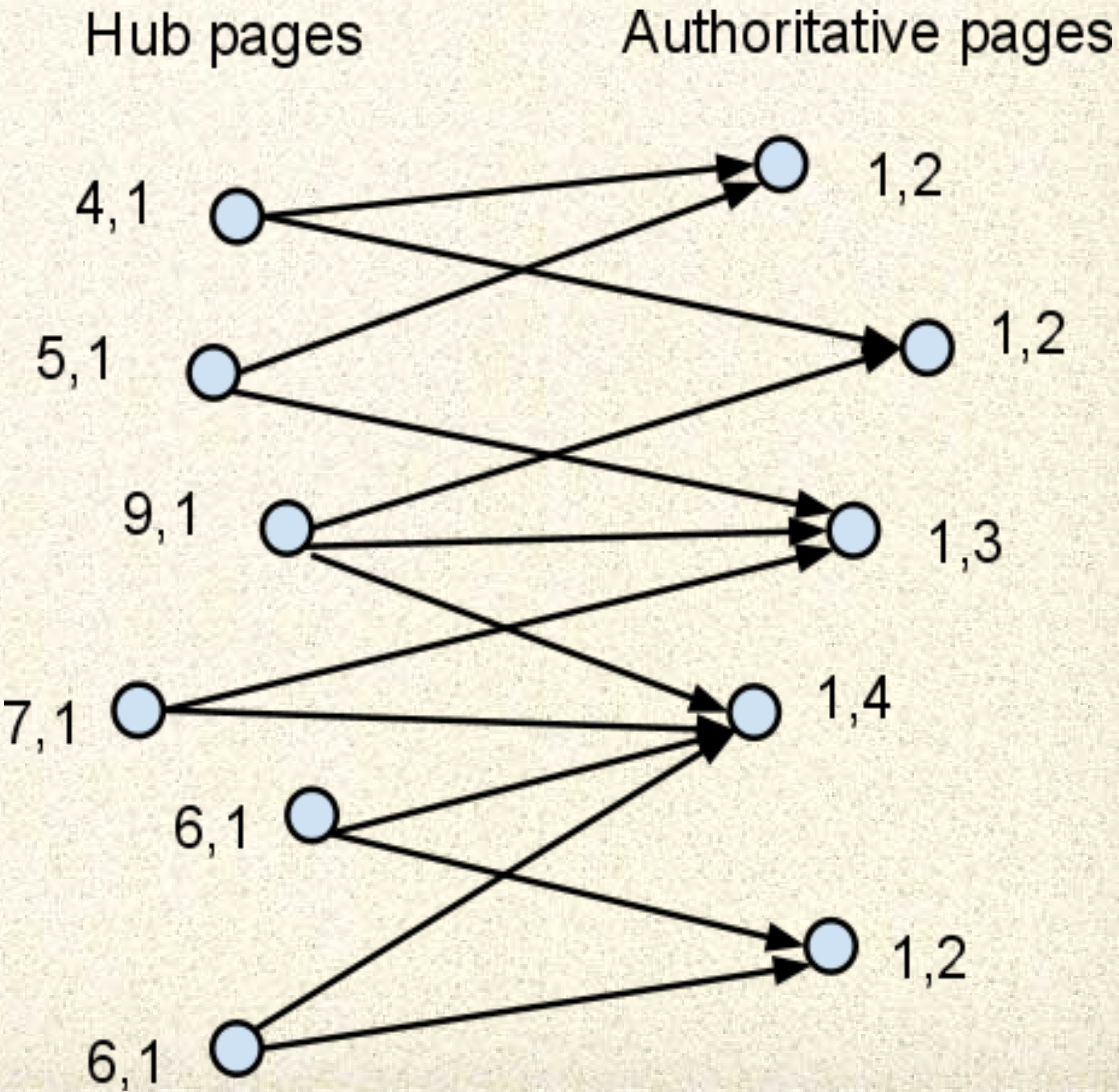
An iterative approach-2

- Update the authority weight by summing the incoming hub weight.



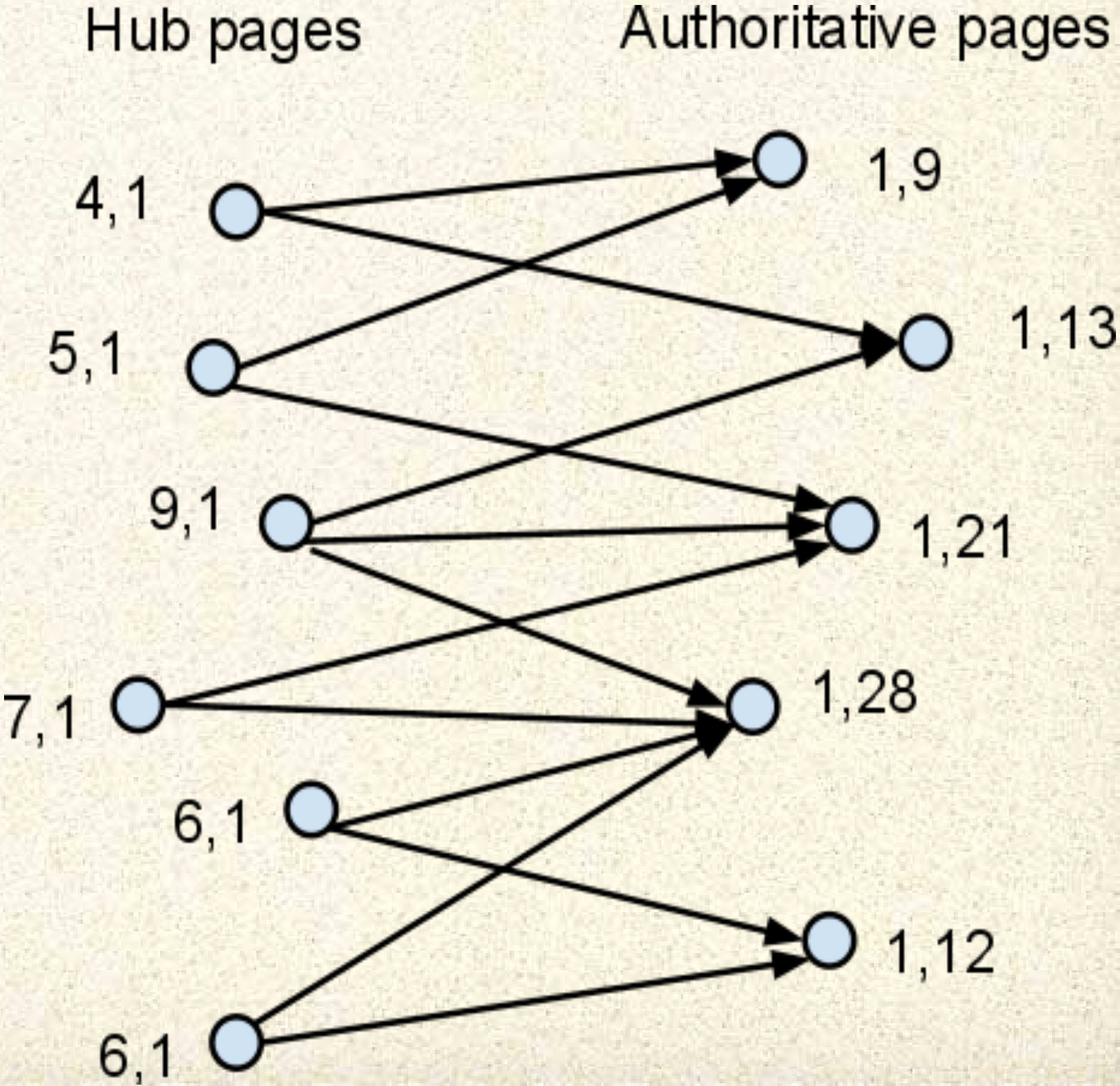
An iterative approach-3

- Update the hub weight by summing the outgoing authority weight.



An iterative approach-4

- Again update the authoritative weight base on the incoming hub weight.



An iterative approach-5

- Observation:

1. After a few iterations, the most authoritative pages will have a very large authority weights.
2. The best hub pages will have a very large hub weights.

- If we normalize the weight after each iteration, each weight will become stable eventually.

Recall duality problems

- Duality problems are materialized in trying to identify two sets of inter-related concepts.
- In authoritative pages example, "authoritative page" and "hub page" are the two set of inter-related concepts. This two set of pages are related in the way that they are densely linked together and they can mutual enforce each other.
- We identified them by an iterative approach.

Problem 2: Extract author-title pair

- In this problem, we are interesting to extract the author-title pair of a book from the web, with a small set of author-title pairs given.
- Here, we defined two concept Relation and Pattern.

- Relation: author-title pair. E.g. (Isaac Asimov, The Robots of Dawn).

- Pattern: how author-title pairs appear in a web page. E.g.

<LI><B>*title*</B> by *author*
<i>*title*</i> by *author*

Pattern Relation Duality

- We can construct a very good set of author-title pairs simply by crawling the web and matching to a good set of patterns.
- Given a good set of author-title pairs, we can build a good set of patterns about how those pairs appears on the web.
- The combination of the ability to find author-title pair from patterns and patterns from author-title pair forms the basic of the approach.

Algorithm

1. R' <- Sample

Start with a small sample, R' of the target relation. This sample is given by the user and can be very small.

2. O <- FindOccurrences(R', D)

Find all occurrences of tuples of R' in D. In our example, these were nearby occurrences of the author and the title of a book in text.

3. P <- GenPatterns(O)

Generate patterns based on the set of occurrences. The patterns need to have a low error rate and high coverage.

4. R' <- M_D(P)

Search D for tuples matching any of the patterns.

5. If R' is large enough, return. Else go to step 2.

Run the algorithm-1

- Relations (It is provided by user initially)

* Isaac Asimov	The Robots of Dawn
* David Brin	Startide Rising
* James Gleick	Chaos: Making a New Science

- Patterns

Empty

Run the algorithm-2

- Relations →

* Isaac Asimov	The Robots of Dawn
* David Brin	Startide Rising
* James Gleick	Chaos: Making a New Science

- Patterns

* <LI><B>*title*</B> by *author*

* <i>*title*</i> by *author*

* *author* || *title* ||

Run the algorithm-3

- Relations

* Isaac Asimov	The Robots of Dawn
* David Brin	Startide Rising
* James Gleick	Chaos: Making a New Science
* H.D. Everett	The Death-Mask and Other Ghosts
* H.G. Wells	First Men in the Moon
* H. G. Wells	Science Fiction: Volume 2

.....

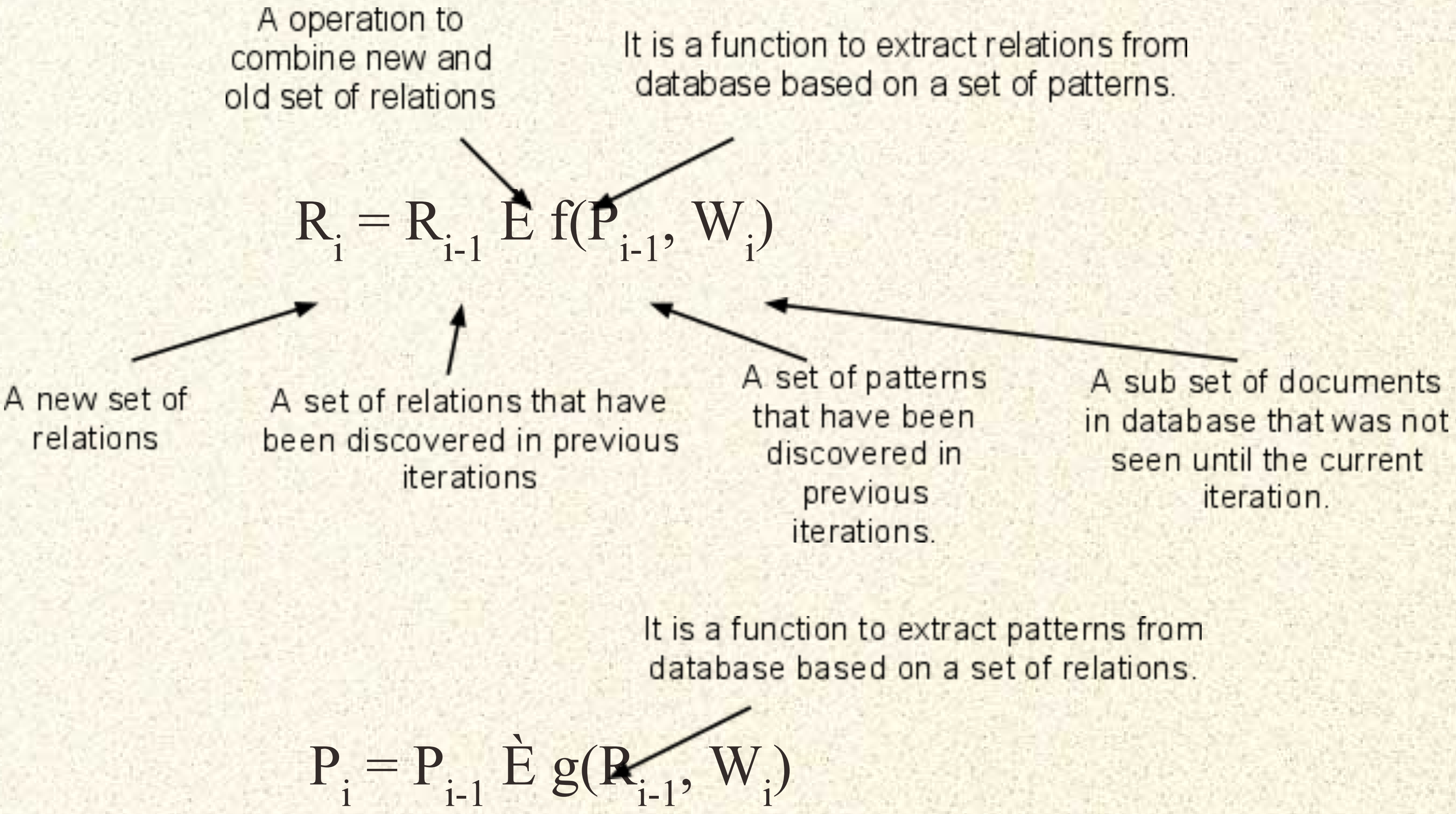
- Patterns

* <LI><B>_title_</B> by _author_

* <i>_title_</i> by _author_

* _author_ || _title_ ||

Formalize duality of relations and patterns



A operation to combine new and old set of relations

It is a function to extract relations from database based on a set of patterns.

$R_i = R_{i-1}$ È $f(P_{i-1}, W_i)$

A new set of relations

A set of relations that have been discovered in previous iterations

A set of patterns that have been discovered in previous iterations.

A sub set of documents in database that was not seen until the current iteration.

It is a function to extract patterns from database based on a set of relations.

$P_i = P_{i-1}$ È $g(R_{i-1}, W_i)$

Higher level duality problems

- The problems before are 1-level duality problem.
- 2-level duality problem is defined as followings:

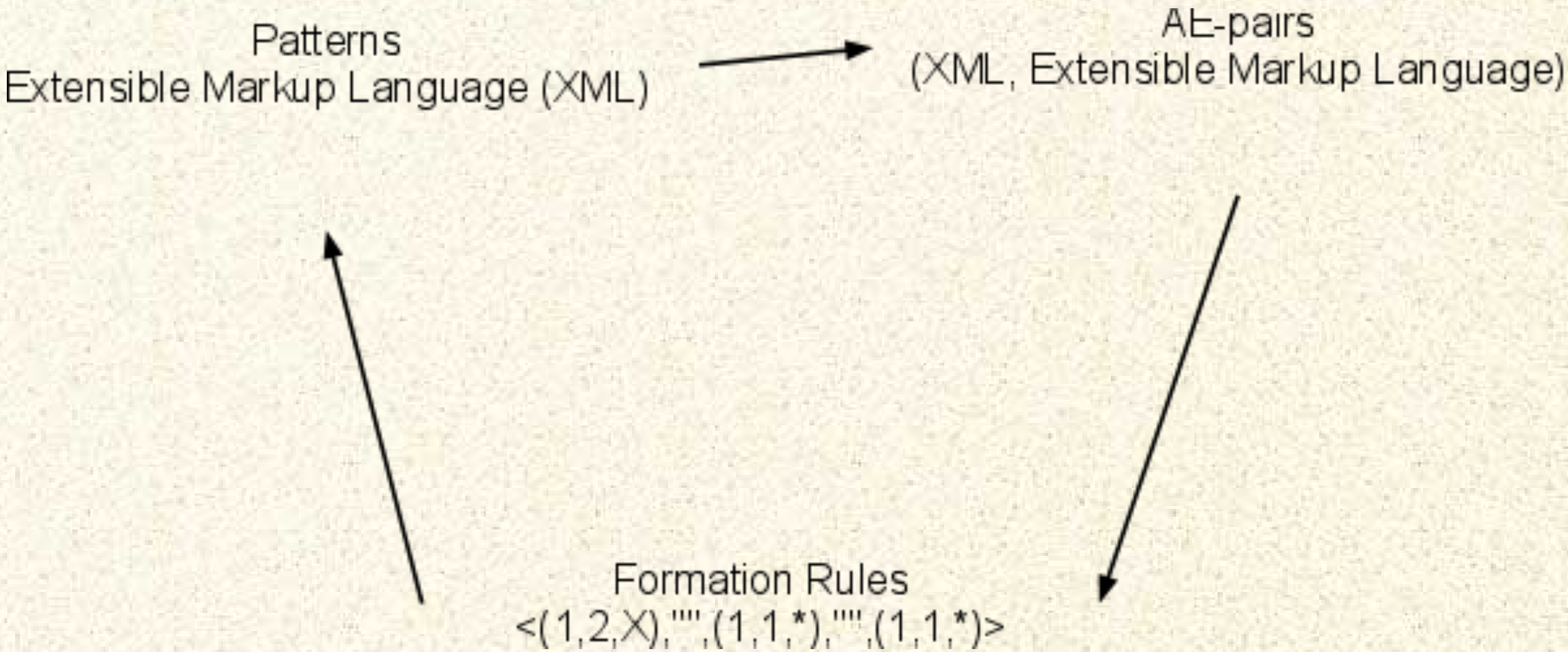
  $R_i = R_{i-1}$ È $f(P_{i-1}, W_i)$
  $P_i = P_{i-1}$ È $g(S_{i-1}, W_i)$
  $S_i = S_{i-1}$ È $h(R_{i-1}, W_i)$

- It means that an approximation of R in a particular iteration may depend on an approximation to P in a previous iteration, which in turn may depend on an approximation to S in a previous iteration.

2-level duality problem

- Problem: We want to identify acronyms and their expansions in the WWW. E.g. (XML, extensible markup language).
- In order to identify acronym-expansion-pairs (AE-pairs), we need to identify the patterns AE-pairs appears on the web.
- In order to identify the pattern, we need to find out a set of formation rules, which states the way how AE-pairs are formed.

2-level duality problem



Patterns
Extensible Markup Language (XML)

AE-pairs
(XML, Extensible Markup Language)

Formation Rules
<(1,2,X),"",(1,1,*),"",(1,1,*)>

Algorithm

```
1. initial set of AE-pairs: $R_0$ (provided by user)
   initial set of patterns: $P_0$
   initial set of formation rules: $S_0$
2. Set i = 1
3. Let $W_i$ be a set of new web pages crawled.
       $R_i = R_{i-1}$ È $f(P_{i-1}, W_i)$
       $S_i = S_{i-1}$ È $h(R_{i-1})$
       $P_i = P_{i-1}$ È $g(R_{i-1}, S_{i-1}, W_i)$
4. Set i = i+1
5. If steady state, stop, otherwise go to step 3.
```

Conclusion

- This paper explore the duality problem of how entities are related on the web.
- This paper formalized the iterative process of mining for patterns and relations over text, structures, and links.
- Given that the web is a great source of information where information itself is buried under loosely defined structures, mining relations and patterns is an efficient way to discover information.

Question?