Example Link Analysis applications in CiteSeer

Compiled by Cornelia Caragea, Sujatha Das

August 22, 2014

History of link analysis

- Bibliometrics
  - Citation analysis since the 1960’s
  - Citation links to and from documents
    - Basis of pagerank idea

Bibliometric techniques use citation analysis to measure the similarity of journal articles or their importance

Bibliometrics

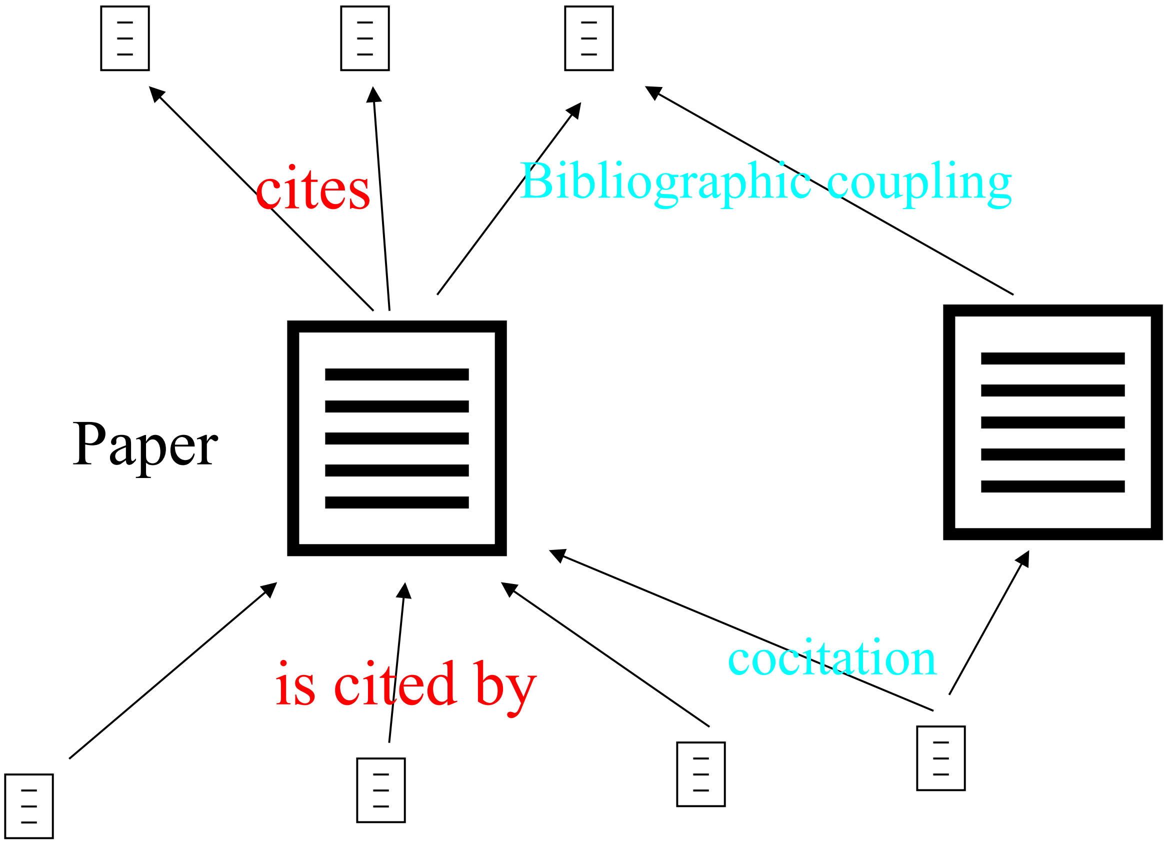
Bibliographic coupling: two papers that cite many of the same papers

Cocitation: two papers that were cited by many of the same papers

Impact factor (of a journal): frequency with which the average article in a journal has been cited in a particular year or period

Citation frequency

Citation Graph



Note that academic citations nearly always refer to earlier work.

Citations in Documents

- Many standard documents include *bibliographies* (or *references*), explicit *citations* to other previously published documents.

- Using citations as links, standard corpora can be viewed as a graph.
- The structure of this graph, independent of content, provides interesting information about the similarity of documents and the structure of information.

Citations vs. Web Links

- Web links are a bit different than citations:
  - Many links are navigational.
  - Many pages with high in-degree are portals not content providers.
  - Not all links are endorsements.
  - Company websites don’t point to their competitors.
  - Citations to relevant literature is enforced by peer-review.

Using Citation Links in CiteSeer

- Several networks (author-document, co-authorship, citation network)
- Several applications of citation links (similar to hyperlinks) and citation contexts (similar to anchor text)
  - Classifying a cited paper (extends, refutes, confirms, credits, applies)
  - Finding communities of authors, influential authors
  - Extracting keyphrases
  - Extracting topical trends over time

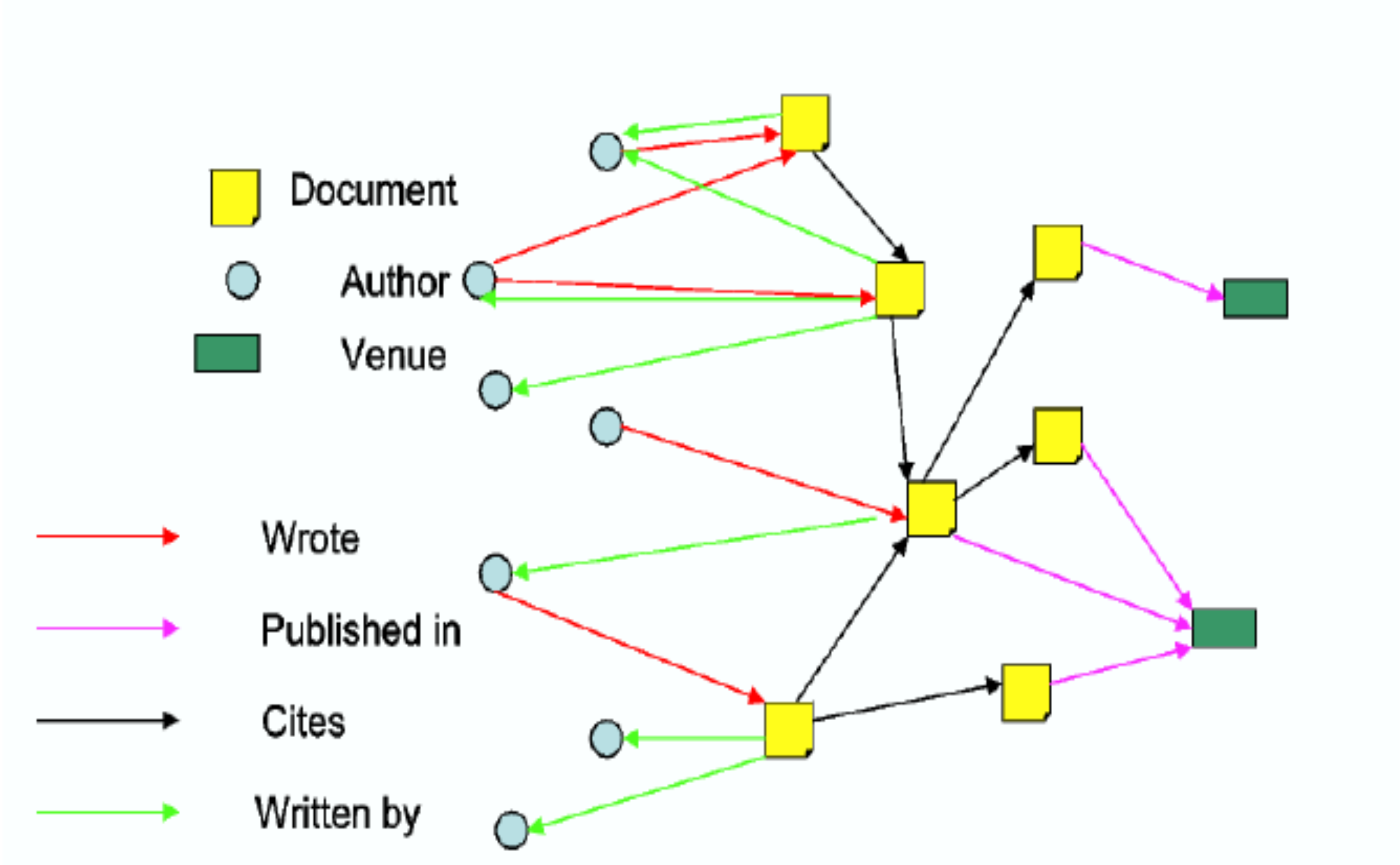
Enabling Expert Search in CiteSeer

- Expert Search is a common task
  - TREC expert-finding task (2005)
    - Goal: Find experts in an enterprise who can answer the "query”
  - In academic/digital library setting
    - Find experts on a particular topic (e.g. "entity ranking in graphs", "machine learning techniques for object recognition”)
    - Evidence of expertise
      - Authored documents, homepages, citation graph, venues of publications etc.

Extending PageRank for Expert Search

- Basic Idea:
  - An initial set of document nodes is activated based on match with the query.
  - Use PageRank on the subgraph to combine both structural and query-dependent aspects into a single ranking model.
  - Efficient techniques exist for computing scores on large, sparse graphs
  - Different types of edges are possible (author-document, citation, author-homepage) but simple extensions of PageRank are possible.

Figure: Sample typed graph



Why Keyphrase Extraction?

- Large number of scholarly documents on the Web
  - Keyphrases allow for *efficient processing of more information in less time*.
- The “concepts” in documents are not always directly available
  - Need to be gleaned from the multitude of details in documents.
  - Keyphrases are useful in many applications such as topic tracking, information filtering and search.

Examples of Keyphrases: A snippet from the 2010 best paper award winner in the WWW conference

Factorizing Personalized Markov Chains for Next-Basket Recommendation

by Rendle, Freudenthaler, and Schmidt-Thieme

"Recommender systems are an important component of many websites. Two of the most popular approaches are based on matrix factorization (MF) and Markov chains (MC). MF methods learn the general taste of a user by factorizing the matrix over observed user-item preferences. *[…]* In this paper, we present a method bringing both approaches together. Our method is based on personalized transition graphs over underlying Markov chains. *[…]* We show that our factorized personalized MC (FPMC) model subsumes both a common Markov chain and the normal matrix factorization model. For learning the model parameters, we introduce an adaption of the Bayesian Personalized Ranking (BPR) framework for sequential basket data. *[…]*"

- Keyphrase extraction is the task of automatically extracting descriptive phrases or concepts from a document.

From Data to Knowledge

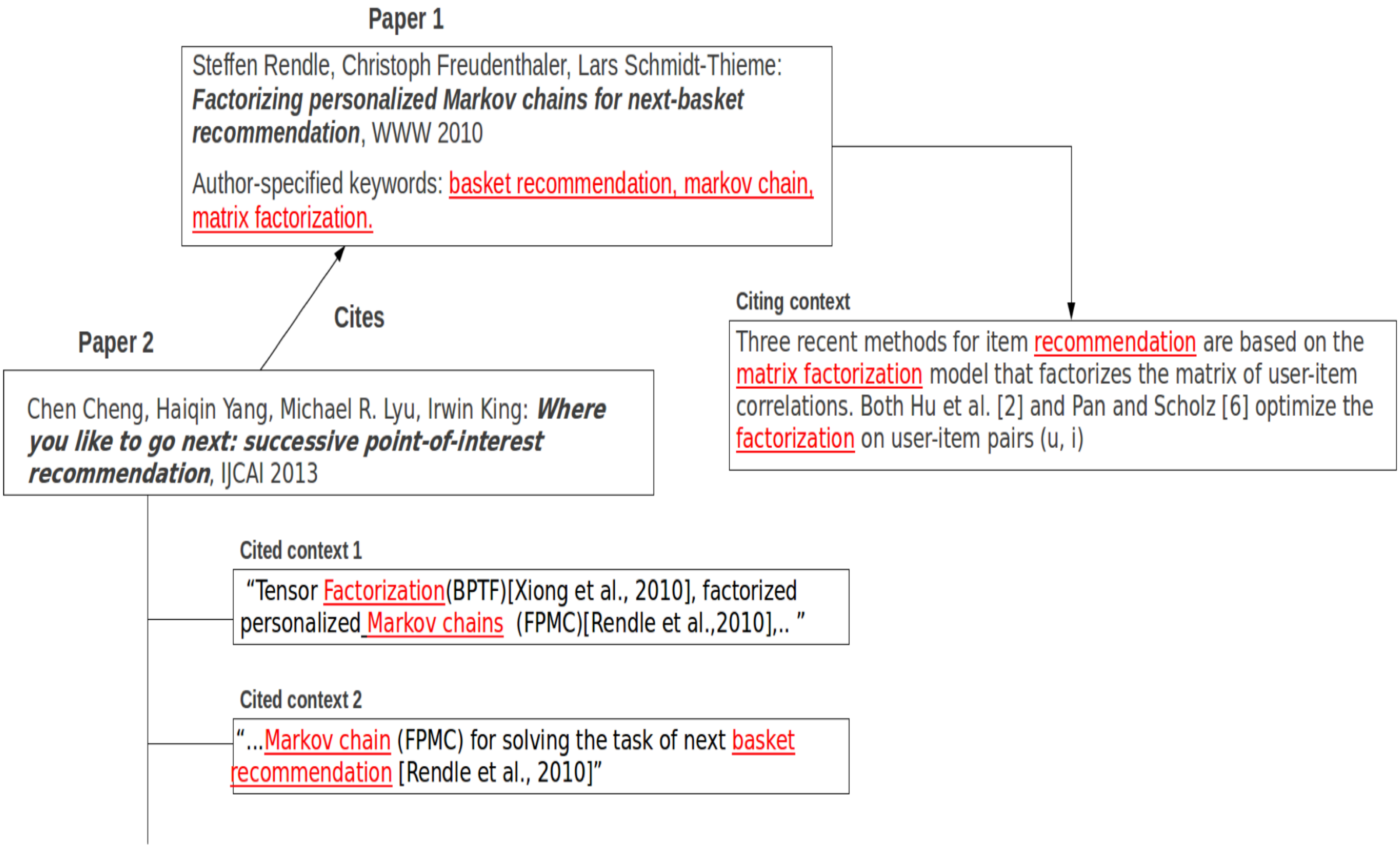
A typical scientific research paper:

- Proposes new problems or extends the state-of-the-art for existing research problems
- Cites relevant, previously-published research papers in appropriate *contexts*.

The citations between research papers gives rise to an interlinked document network, commonly referred to as the *citation network*.

Citation Networks

- In a citation network, information flows from one paper to another via the citation relation (Shi et al, 2010)
- Citation contexts capture the influence of one paper on another as well as the flow of information
- Citation contexts or the short text segments surrounding a paper's mention serve as “micro summaries” of a cited paper!



Paper 1

Steffen Rendle, Christoph Freudenthaler, Lars Schmidt-Thieme: ***Factorizing personalized Markov chains for next-basket recommendation***, WWW 2010

Author-specified keywords: basket recommendation, markov chain, matrix factorization.

Citing context

Three recent methods for item recommendation are based on the matrix factorization model that factorizes the matrix of user-item correlations. Both Hu et al. [2] and Pan and Scholz [6] optimize the factorization on user-item pairs (u, i)

Paper 2

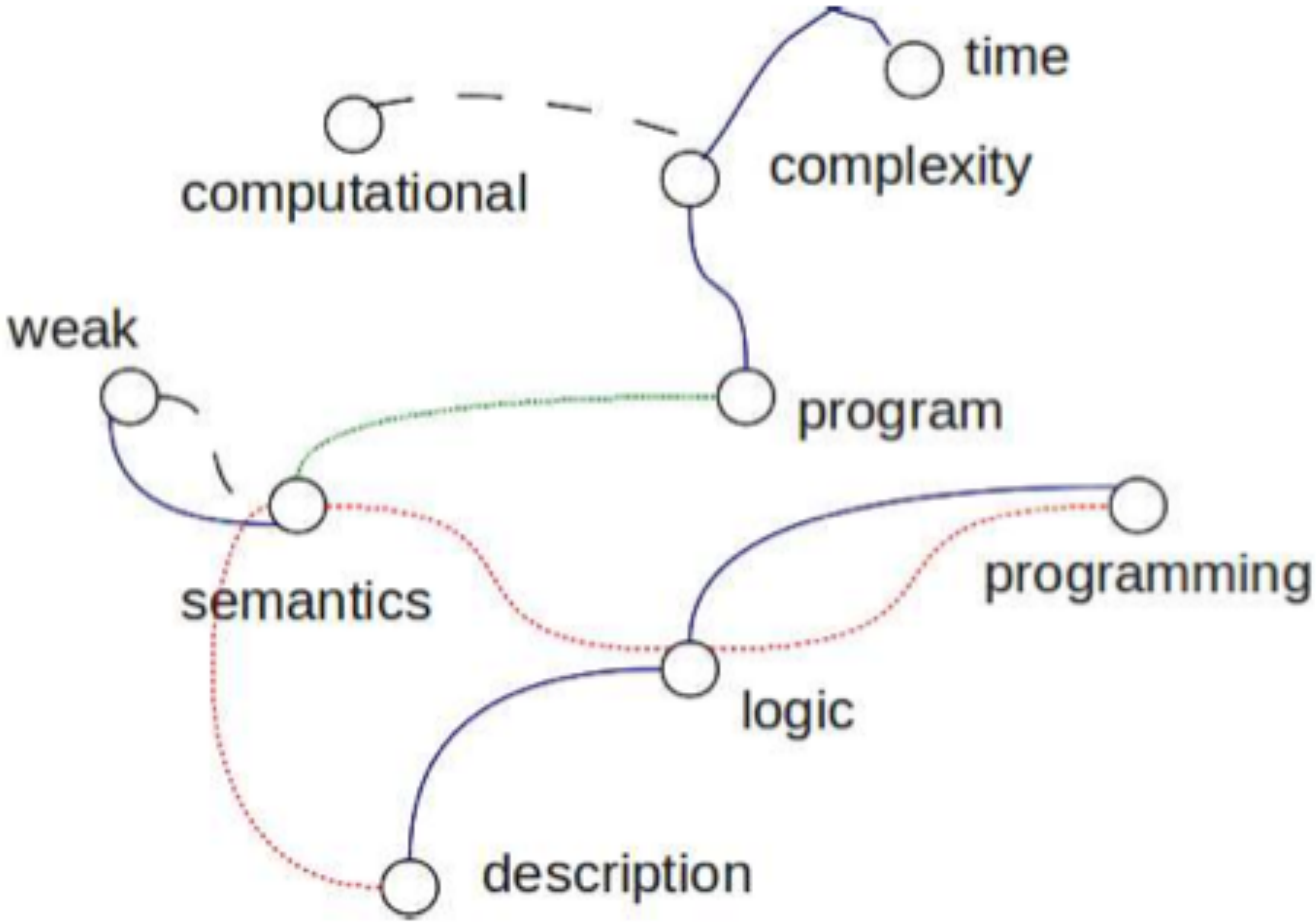
Chen Cheng, Haiqin Yang, Michael R. Lyu, Irwin King: ***Where you like to go next: successive point-of-interest recommendation***, IJCAI 2013

Cited context 1

"Tensor Factorization(BPTF)[Xiong et al., 2010], factorized personalized Markov chains (FPMC)[Rendle et al.,2010],.. "

Cited context 2

"...Markov chain (FPMC) for solving the task of next basket recommendation [Rendle et al., 2010]"



CiteTextRank incorporates information from document content as well as *citation contexts* while scoring candidate words for keyphrase extraction.

Conclusions

- Citation links and contexts are very important in CiteSeer, the flow of information via these links can effectively improve several tasks such as expert search and keyphrase extraction
- Link analysis of various networks in CiteSeer are useful in understanding various phenomena such as topical trends and evolution, influential authors, and author communities

References

1. Sujatha Das Gollapalli, Prasenjit Mitra, C. Lee Giles: Ranking authors in digital libraries. JCDL 2011
2. Sujatha Das Gollapalli, Cornelia Caragea: Extracting Keyphrases from Research Papers Using Citation Networks. AAAI 2014