Measuring Research Quality Using Semantic-based Computational Methods

Gianluca Demartini, Elaine Toms

Information School

Research Quality

- Measuring research quality is important
  - REF
- It is difficult
  - Subjective: 3 stars? 4 stars?
  - Discipline-specific

- Goal: An objective measurement of research quality, automatically computed

Objective

- Measure research impact in a *quantitative* way
  - Use Text Mining approaches
  - At scale
- Obtain estimate indicators of research output quality

Proposed Solution

- Application of **data mining** techniques
- To digital repositories of scientific literature

- Build a software
- Extract **scientific concepts** from publications
- Measure academic impact of research

Approach Overview

- Step 1: Extract scientific concepts from published research
  - Roman Prokofyev, Gianluca Demartini, and Philippe Cudré-Mauroux. Effective Named Entity Recognition for Idiosyncratic Web Collections. In: 23rd International Conference on World Wide Web (WWW 2014).
- Step 2: Create a network of scientific concepts based on the citation graph
- Step 3: Define measure of research quality based on research work impact

Step 1: Scientific Concepts

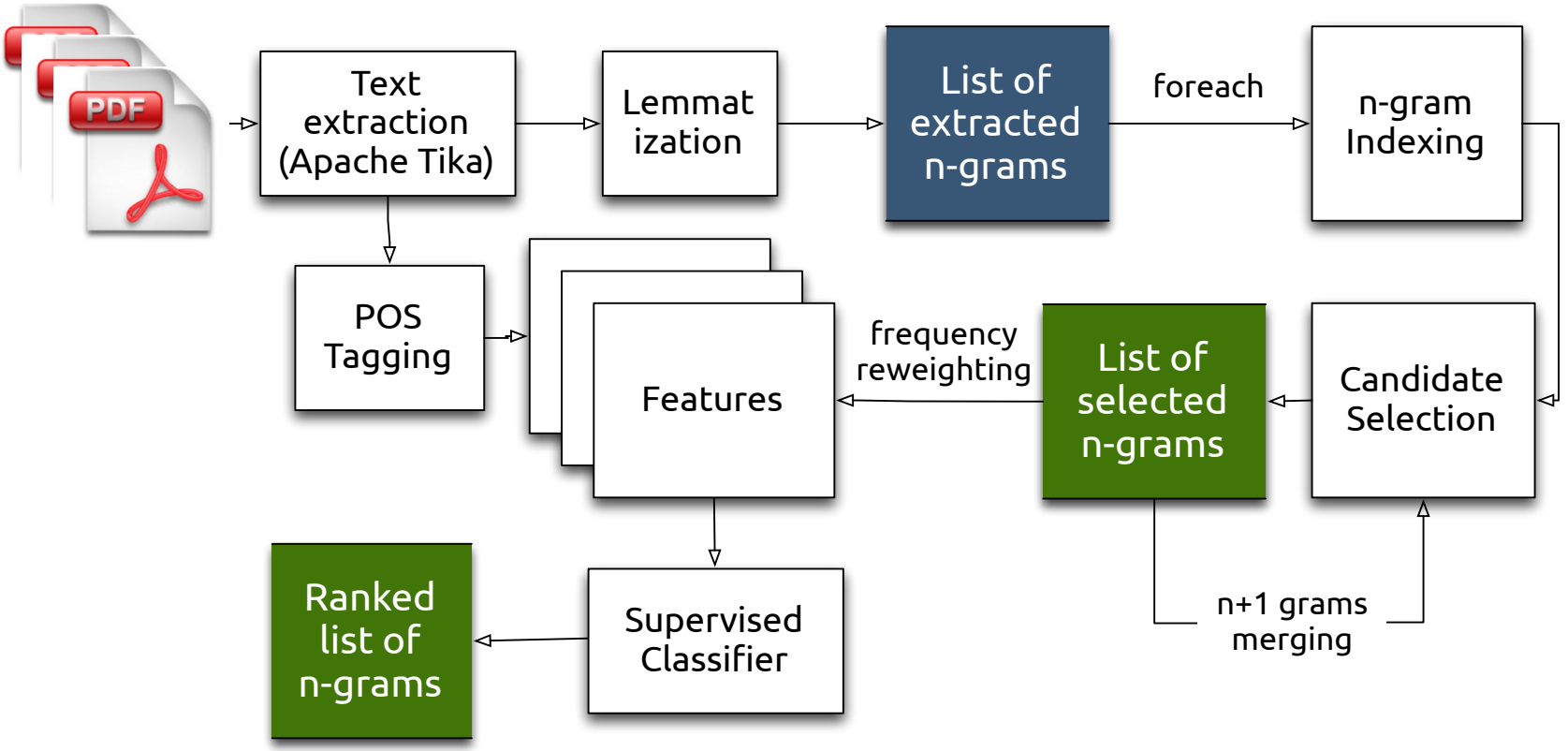
1. INTRODUCTION

Nowadays, accessing information on the Internet through search engines has become a fundamental life activity. Current web search engines usually provide a ranked list of URLs to answer a query. This type of information access does a good job for dealing with simple navigational queries by leading users to specific websites. However, it is becoming increasingly insufficient for queries with vague or complex information need. Many queries serve just as the start of an exploration of related information space. Users may want to know about a topic from multiple aspects. Organizing the web content relevant to a query according to user intents would benefit user exploration. In addition, a list of URLs couldn't directly satisfy user information need. Users have

- search engine
- web search engine
- navigational query
- user intent
- information need
- web content
- …

Entity type: scientific concept

Extraction Pipeline



From: Roman Prokofyev, Gianluca Demartini, and Philippe Cudré-Mauroux. Effective Named Entity Recognition for Idiosyncratic Web Collections. In: 23rd International Conference on World Wide Web (WWW 2014).

A Citation Network

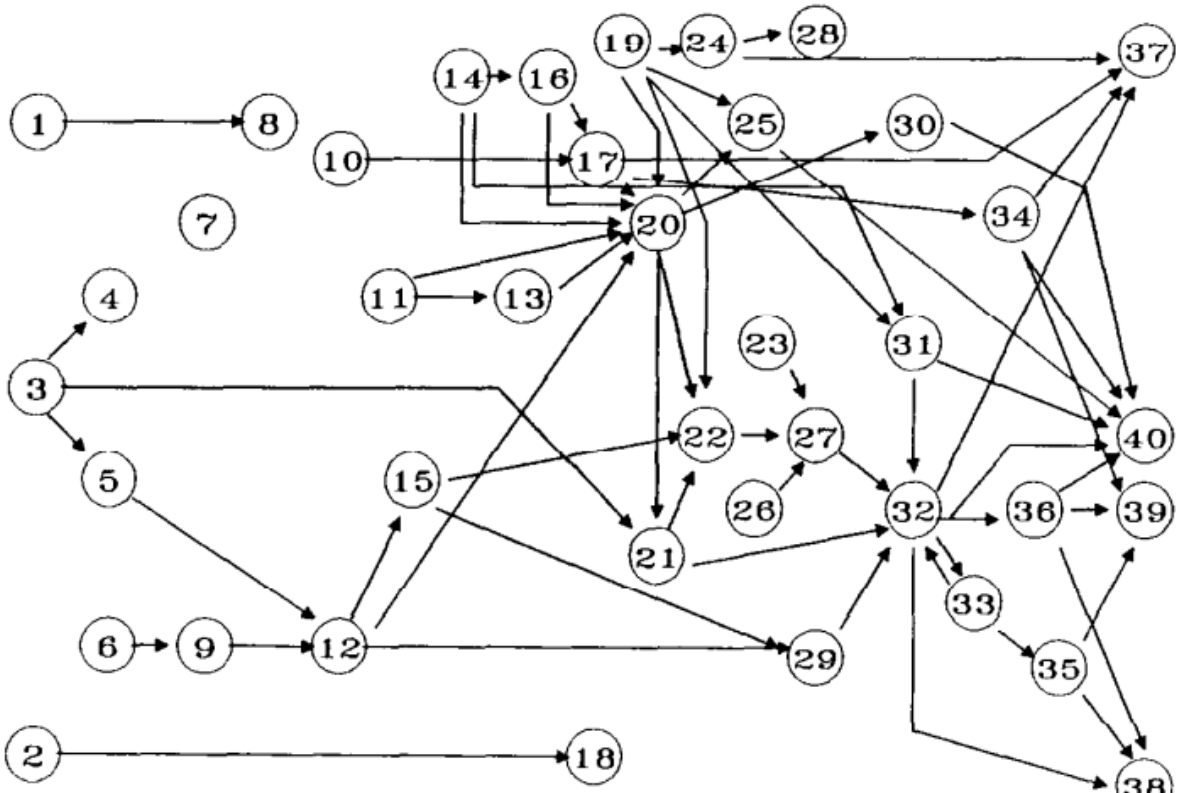
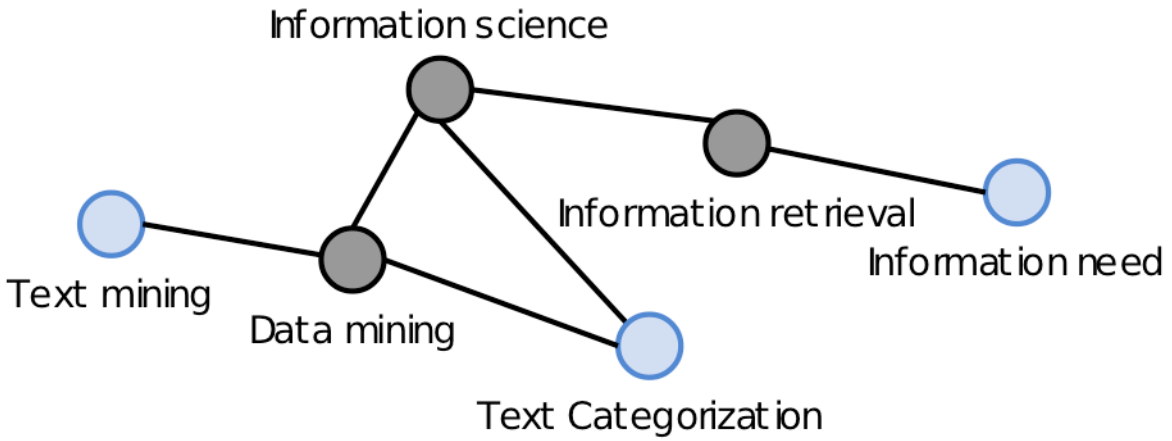


Fig. 1. DNA theory network.

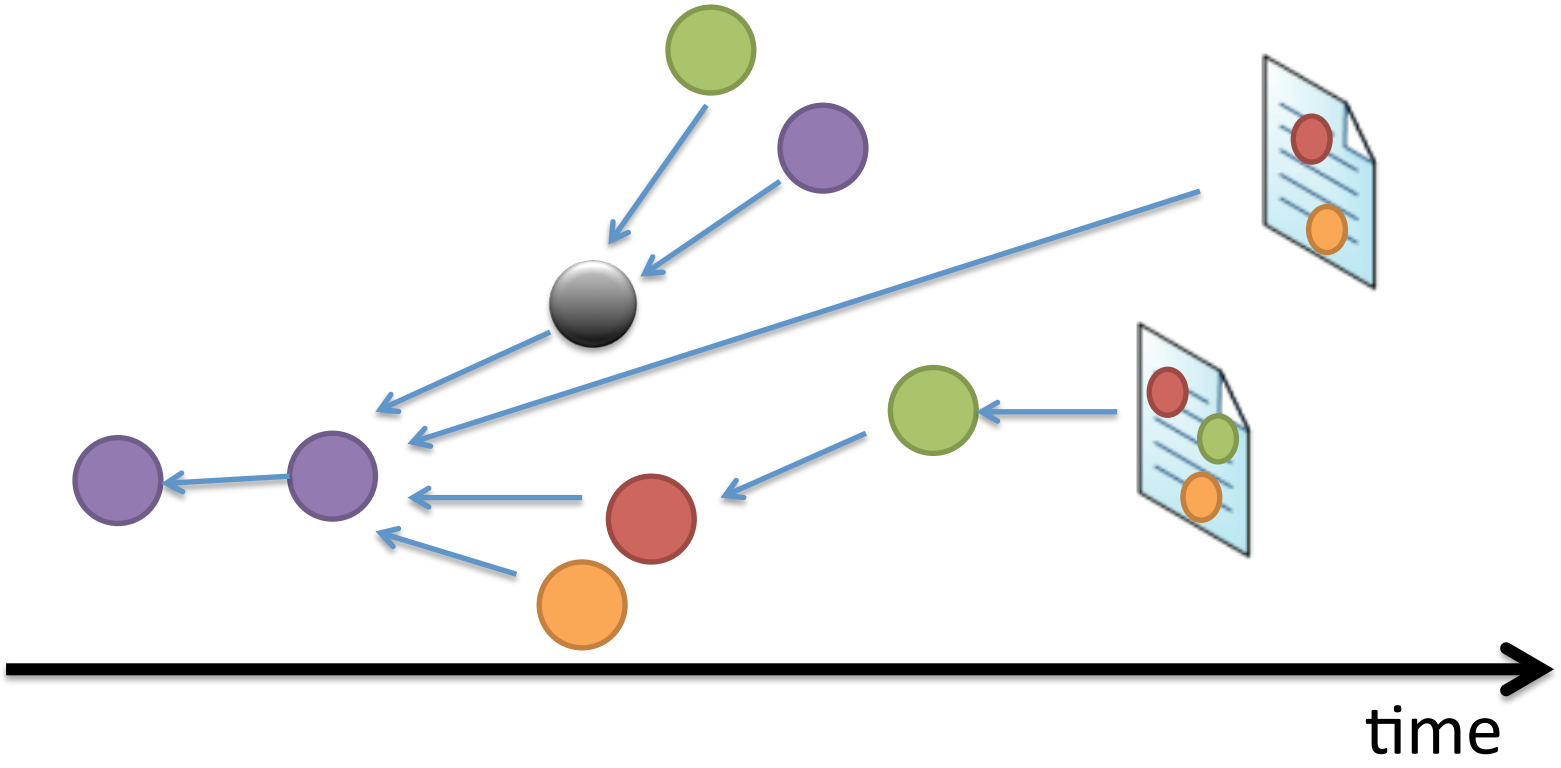
From "Connectivity in a citation network: The development of DNA theory" Norman P. Hummon, Patrick Dereian. In: Social Networks vol. 11, no. 1, pp. 39-63, 1989

Step 2: A Scientific Concept Network

- Based on the citation network
- A graph inter-connecting scientific concepts



Evolving Concept Networks



Applications of Evolving Concept Networks

- We can, automatically
  - Measure scientific impact at the concept level
  - Understand how scientific concepts propagate within a discipline
  - Use Co-occurrence and relatedness of concepts
  - Leverage the **textual context** in which
    - Citations are used
    - Scientific concepts are referred to

New Research Reputation Measures

- Going beyond *h-index*, *impact factor,* and related measures
- Looking at the actual scientific impact in a community rather than counting citations

Conclusions

- Measuring scientific quality and impact is key
- Go from the ***document*** to the ***concept*** level
- Leverage the concept networks and citation context
- Automatize research evaluation to make it more objective and less expensive

- Next: Use our **concept extraction** system to build **concept networks** and **impact measures** at scale
  - Microsoft released 80+M paper dataset