BIBLIOMETRICS-AIDED RETRIEVAL

Wolfgang Glänzel

Centre for R&D Monitoring and Dept. MSI, KU Leuven, Belgium

ECOOM

Structure of presentation

Information Retrieval for Bibliometrics?

Application of bibliometrics has increasingly shifted towards meso and micro studies in the sense of both actors and topic analysis.
One consequence is the necessity of proper subject delineation (domain studies, interdisciplinary research, emerging topics).
Subject delineation strongly relies on IR methods through complex *search strategies*.

Bibliometrics for Information Retrieval?

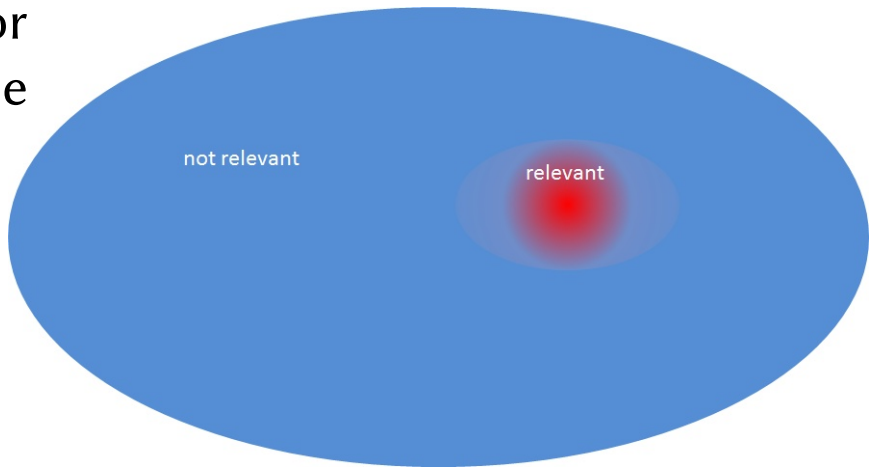
Bibliometrics, in turn, provides important techniques to improve the efficiency of IR. Similarity/distance measures defined on direct citations, bibliographic coupling, lexical relationship and or even "core documents" can facilitate and improve the retrieval of scientific information.

Both Bibliometrics and Information Retrieval may serve as mutual input and can be combined in an iterative way. This combination will be shown in the following.

The initial situation

Bibliometrics, in general, requires specific retrieval. The borderline between relevant and not relevant documents is fuzzy and often determined by users or the actors in the domain in question. Sometimes it has to be adjusted according to the actual needs.

The scope of the study decides whether documents in the red or the purple circle are used for the bibliometric analysis.



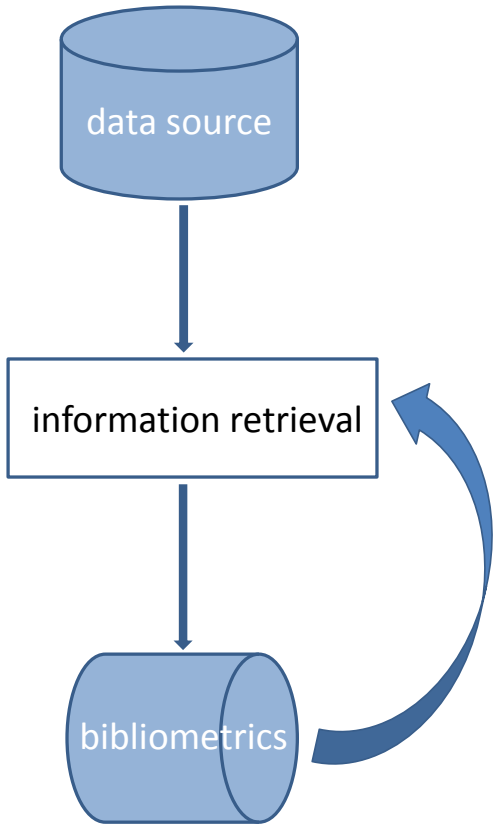
The background

- The objectives of subject delineation in the framework of domain studies essentially differ from the goals of traditional information retrieval.
- In addition, bibliometrics allows including also 'metric' components in the search strategy.
- Thresholds of the strength of citation, bibliographic-coupling or textual links can be used to fine-tune the metric component.

Bibliometrics-aided retrieval is a combination of traditional search strategies with advanced bibliometrics methods.

- Zitt & Bassecaulard, *IPM*, 2006
- Glänzel et al., *STI Conference*, 2006

How to combine Information Retrieval and bibliometrics?



When Bibliometrics meets Information Retrieval ...

The first step comprises so-called *unconditional* criteria: $(UC_1, \ldots, UC_k)$ with $k \geq 0$.

In a second step this set is extended by potentially relevant documents. This includes so-called *conditional* criteria $\{CC_1, \ldots, CC_m, \ldots, CC_{m+n}\}$ with $m, n > 0$ or $m = n = 0$. Assume that we have at least one conditional criterion and another conditional or unconditional criterion.

The *bibliometric retrieval* (BR) strategy can then be defined as the following logical combination

$$BR = (UC_1 \vee \ldots \vee UC_k) \vee ((CC_1 \vee \ldots \vee CC_m) \wedge (CC_{m+1} \vee \ldots \vee CC_{m+n}).$$

Example: Bioinformatics (e.g., Glänzel et al., 2006, 2009)

UC1: Journal in WoS = BIOINFORMATICS (formerly COMPUTER APPLICATIONS IN THE BIOSCIENCES), JOURNAL OF COMPUTATIONAL BIOLOGY, BRIEFINGS IN BIOINFORMATICS, BMC BIOINFORMATICS

UC2: Journal in Medline= IN SILICO BIOLOGY, PSB ON-LINE PROCEEDINGS, APPLIED BIOINFORMATICS, PLOS COMPUTATIONAL BIOLOGY

CC1: Keywords in title = BIOINFORMATICS, COMPUTATIONAL BIOLOG*, SYSTEMS BIOLOGY

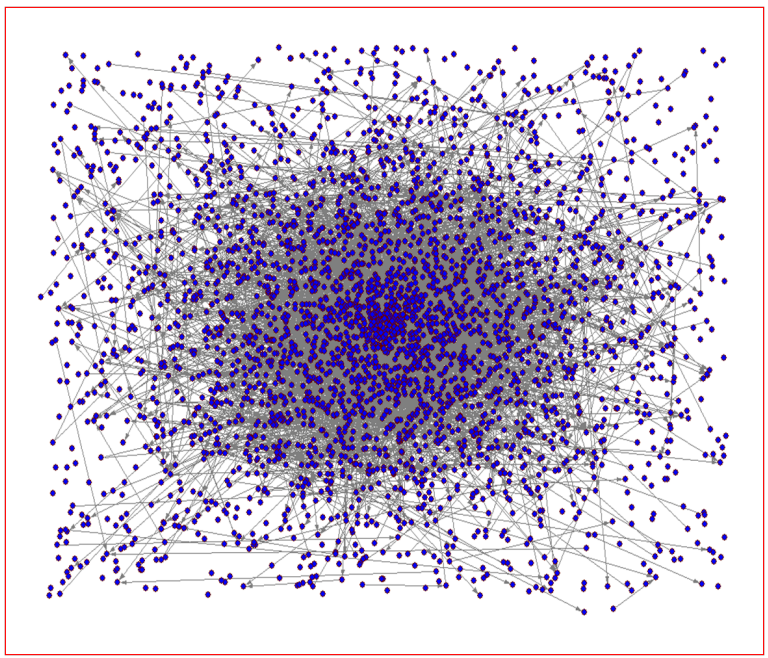
CC2: Related records of UC1

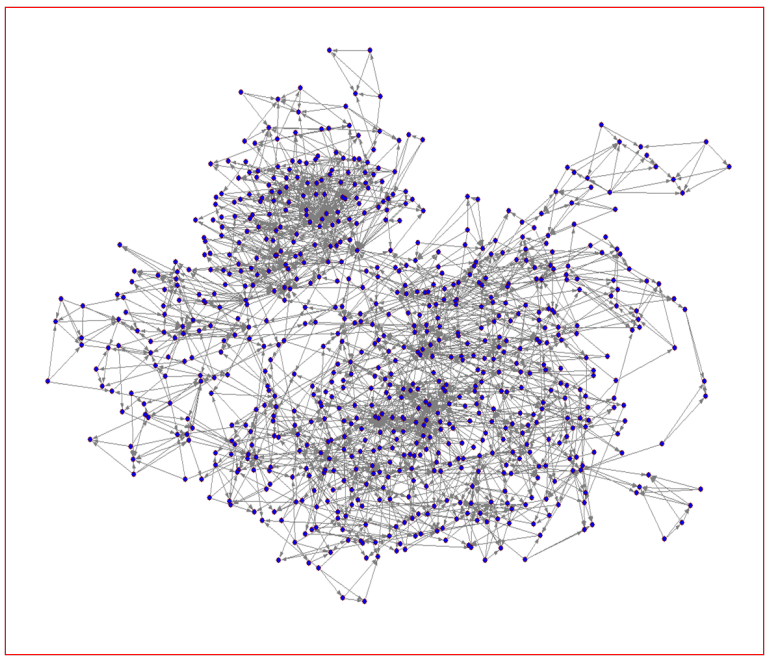
CC3: Cited or citing source of UC1

The search strategy: $BR := (UC_1 \vee UC_2) \vee (CC_1 \wedge (CC_2 \vee CC_3))$

Bibliometrics-aided retrieval

The core set (UC_1 – left) and of all retrieved papers (BR – right) with Kamada-Kawai layout (Glänzel et al., 2006)





Data source: Thomson Reuters Web of Knowledge

New research lines

In order to facilitate the retrieval, especially within rather small areas, BR can be extended by using hybrid similarities. Instead of the combination of citation links or "related records" (based on bibliographic coupling), similarities based on hybrid textual-citation methods can be applied to some of the conditional criteria. This might help avoid too many steps in the logical BR algorithm.

Example

A document is considered relevant if it meets some conditional criterion (CC_j), and is strongly linked based on a hybrid similarity measure to at least a certain number of documents meeting an unconditional criterion (UC_i).

Core documents in bibliometrics-aided retrieval

The notion of a "core" of literature goes back to *co-citation analysis*.

- Small, *JASIS*, 1973.

Definition:

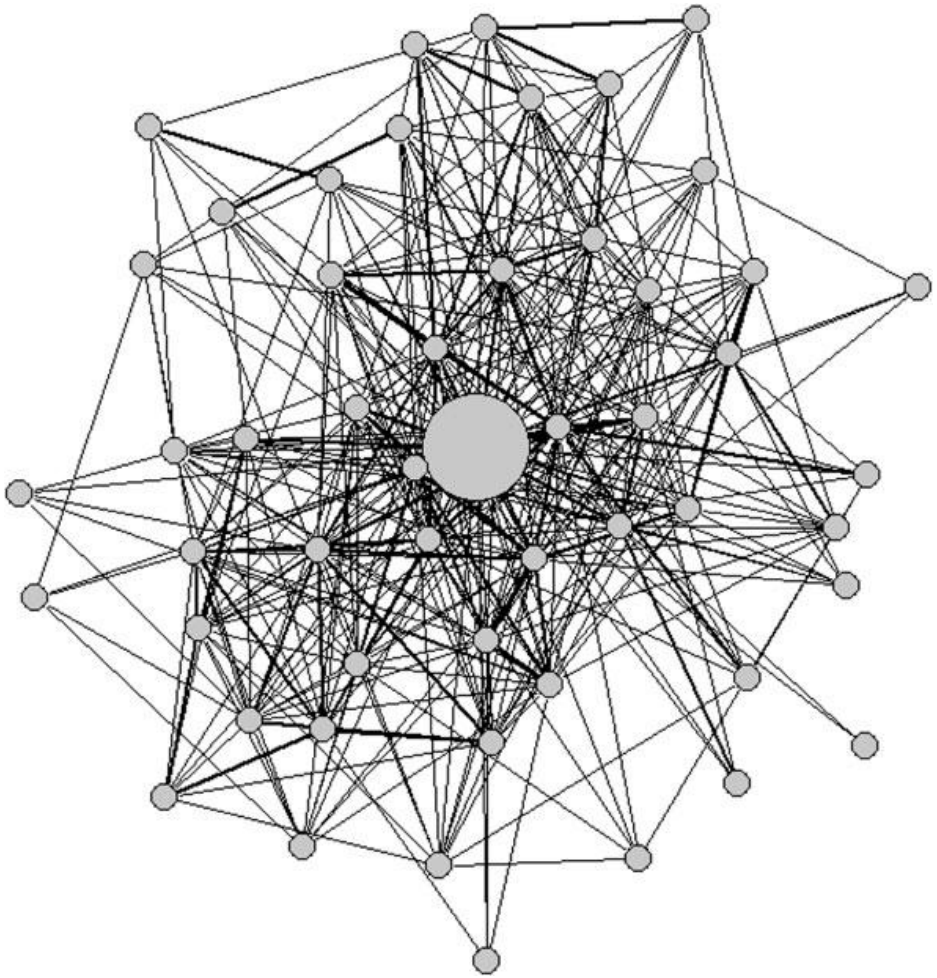
Core documents are defined as papers, which have at least *n* links of at least a given strength *r* according to a given similarity measure.

- Glänzel & Czerwon, *Scientometrics*, 1996
- Glänzel & Thijs, *Scientometrics*, 2012

Core documents can directly be used for data retrieval, namely to identify further relevant documents by following their strong and medium-strong links.

Cluster representation for dynamic analysis

Visualisation of the link environment of a 'core document'
(according to Glänzel & Thijs, 2012)



Data source: Thomson Reuters Web of Knowledge

Conclusions

The fields of applications of the described methods are manifold. Below we give some examples.

- Bibliometrics-aided retrieval is a powerful tool to develop and adjust search strategy at any level of aggregation. It improves even the delineation of complex and interdisciplinary fields and topics.
- Adjustable hybrid (text/citation-based) techniques allow bibliometrics-aided retrieval even in fields where citations do not play an important role.
- Core documents represent the most interlinked papers in a set. Following their strong *and* weaker links might help retrieve relevant information without formulating search queries.
- Metrics can be used for fine-tuning search strategies and to stop retrieval at any level.

Thank you very much for your attention.

Vielen Dank für Ihre Aufmerksamkeit!

Hartelijk dank voor uw aandacht!

¡Muchísimas gracias por su atención!

Köszönöm szépen a figyelmüket!

Molte grazie per la vostra attenzione.

Muito obrigado pela vossa atenção.