Dissemination Patterns of Technical Knowledge in the IR Industry.

Scientometric Analysis of Citations in IR-related Patents

Dr. Ricardo Eito-Brun
Universidad Carlos III de Madrid

ICIC2013
VIENNA, October 15, 2013

Introduction
Patents as a source of data for IC Assessments

- This communication presents the objectives and preliminary results of an academic project.
- Project main objective is to: establish an *innovation activity model* with guidelines to implement successful innovation and technology transfer practices.
- Project specific objectives are:
  - Identify groups within companies and institutions that are leading innovation in specific areas related to Software development for different industries.
  - Identify best-practices regarding systematic Innovation management.
  - Assess the ROI of their innovation efforts.
  - Link conclusions to IC and innovation assessment Models (InnoSpice).

Introduction
Patents as a source of data for IC Assessments

- Patent analysis is a key component in this strategy:
  - Patents are one of the main outputs of innovation efforts.
  - Patents represent the value of innovation results: something that companies want to protect as a potential source of competitive advantage.
  - Patents embody a significant part of the innovation process:
    - Analysis of opportunities.
    - Comparison with existing innovations.
    - Make explicit the contribution to the actual state of knowledge.
  - Patent-based indicators may be used to assess the results of the innovation processes put in place by the organization (not the only ones, of course).

Introduction
Patents as a source of data for IC Assessments

- Patent analysis also offers interesting data regarding:
  - Consumption of information and
  - Knowledge dissemination patterns.
- Examples:
  - Which academic journals have an impact on innovation?
  - To which extent the research done by academic institutions and universities has visibility in the industry?
  - Which is the impact of basic, academic research on “practical innovation”?
  - Which is the impact of previous research made by other companies, probably competitors?
  - How companies are tracking competitors’ activities?
  - Which are the most influential companies/institutions – regarding innovation - on specific knowledge areas?

Introduction
Patents as a source of data for IC Assessments

- These studies are also valuable to improve our knowledge about the historical evolution of specific disciplines.
- In our example, findings provide a better understanding of the evolution of software technologies for Documentation and Information management:
  - Key players from the industry.
  - Inventors
  - Research ideas and innovations.
  - Life-cycle of specific methods and techniques.
  - Areas where these software-based techniques have been applied

Introduction
Patents as a source of data for IC Assessments

- Project steps:
  - Identify sample knowledge areas or domains.
  - Conduct research to identify “leading companies or research groups”.
  - Complete further assessments (interviews, questionnaires) to collect:
    - Best practices and activities for building an Innovation Activity Model.
    - Data about knowledge consumption patterns.
- Patent citation analysis is planned to be used for these steps:
  - Identify “leading organizations” and groups within these organizations.
  - Identify preliminary data about information consumption and tracking of competitors.

Introduction
Patents as a source of data for IC Assessments

- The analysis of patent citations has well-developed theoretical foundations:
  - MIT book, by Adam Jaffe and Manuel Trajtenberg.
- Application of additional bibliometric techniques and metrics can provide interesting views of the data.
- There are problems, anyway:
  - Availability of source data (not all the patents DB include citations in a format easy to process).
  - Restrictions regarding software-based patents .
  - Not all the inventions are, necessarily, patented.
  - Motivation behind citations in patent documents.

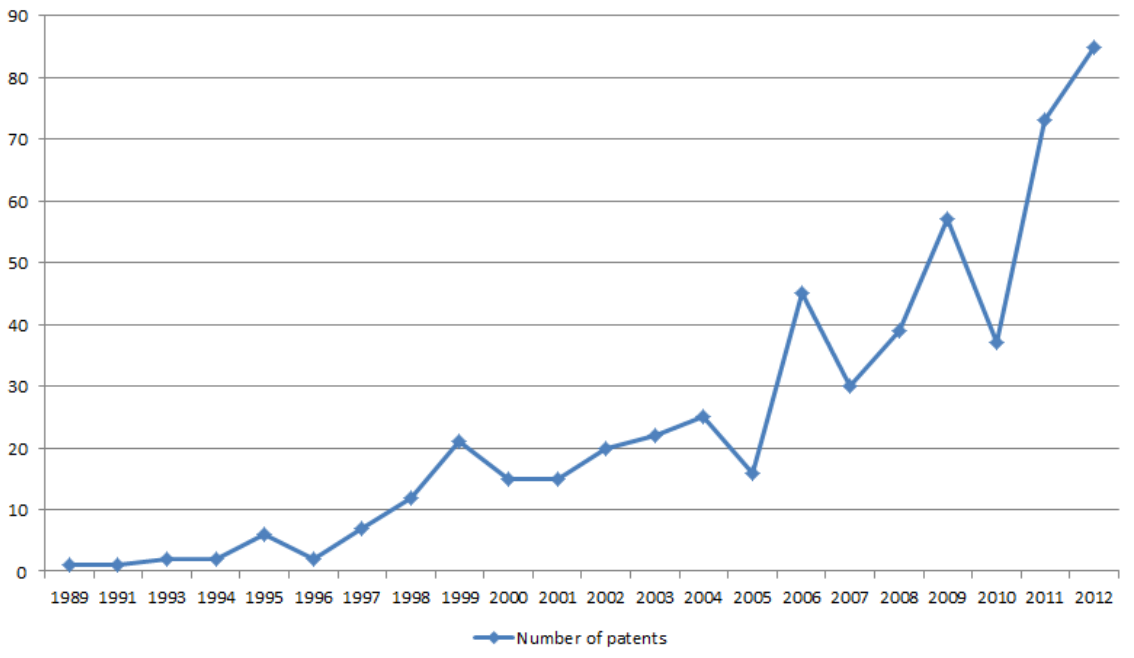
Sample Analysis
Innovations on Text Mining

- The initial scope of the job is focused on Text Mining.
- Text Mining focuses on "*the discovery by computer of new, previously unknown information, by automatically extracting information from different written resources*." (Hearst, 2003)
- Text Mining includes techniques like:
  - Automatic Classification
  - Clustering
  - Information Extraction
  - Text Summarization and automatic abstract generation.
- These techniques share similar theoretical foundations, so in some cases is not easy to assign a contribution to a specific sub-area.
- Preliminary work is done for the "clustering and classification" subset.

Sample Analysis
Innovations on Text Mining

- A preliminary set of patents has been extracted from the Delphion database for "classification and clustering of textual information" (subset of the text mining area).
- Only from the US Patent Office, but not only from US organizations.
- Initial set of 1204 patents.
- Screening of the retrieved patents have restricted the initial set to a sample of 535 patents
- Selected patents include 11884 citations to other patents and 5804 citations to other documents (including patent applications).

Sample Analysis
Innovations on Text Mining. Sample dataset

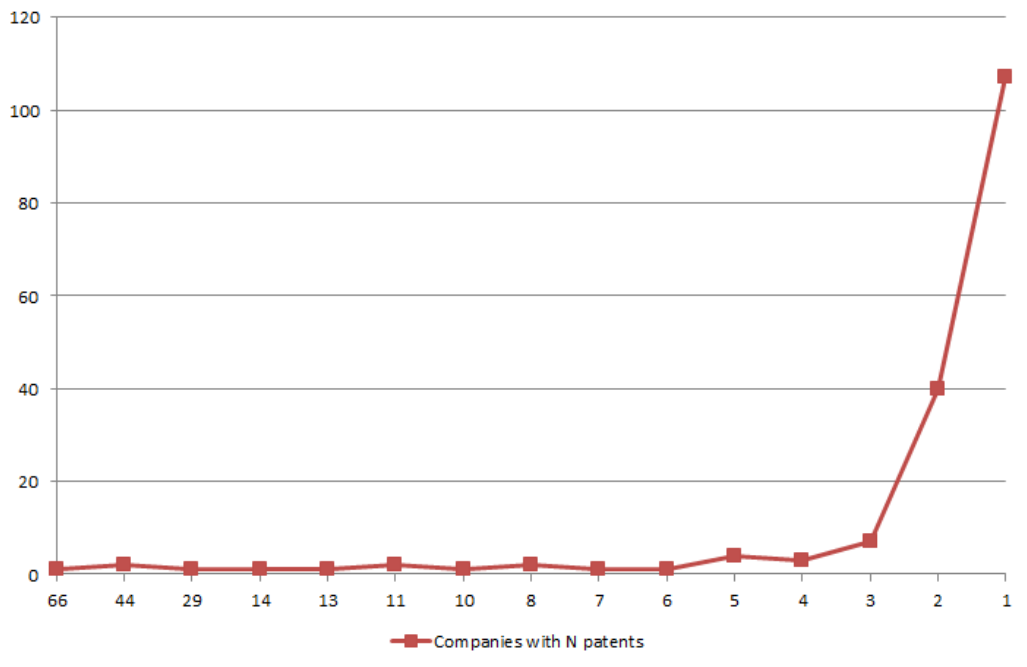


Evolution in number of patents.

Increasing output in the last years

Significant increment in the last 10 years.

Sample Analysis
Innovations on Text Mining. Sample dataset



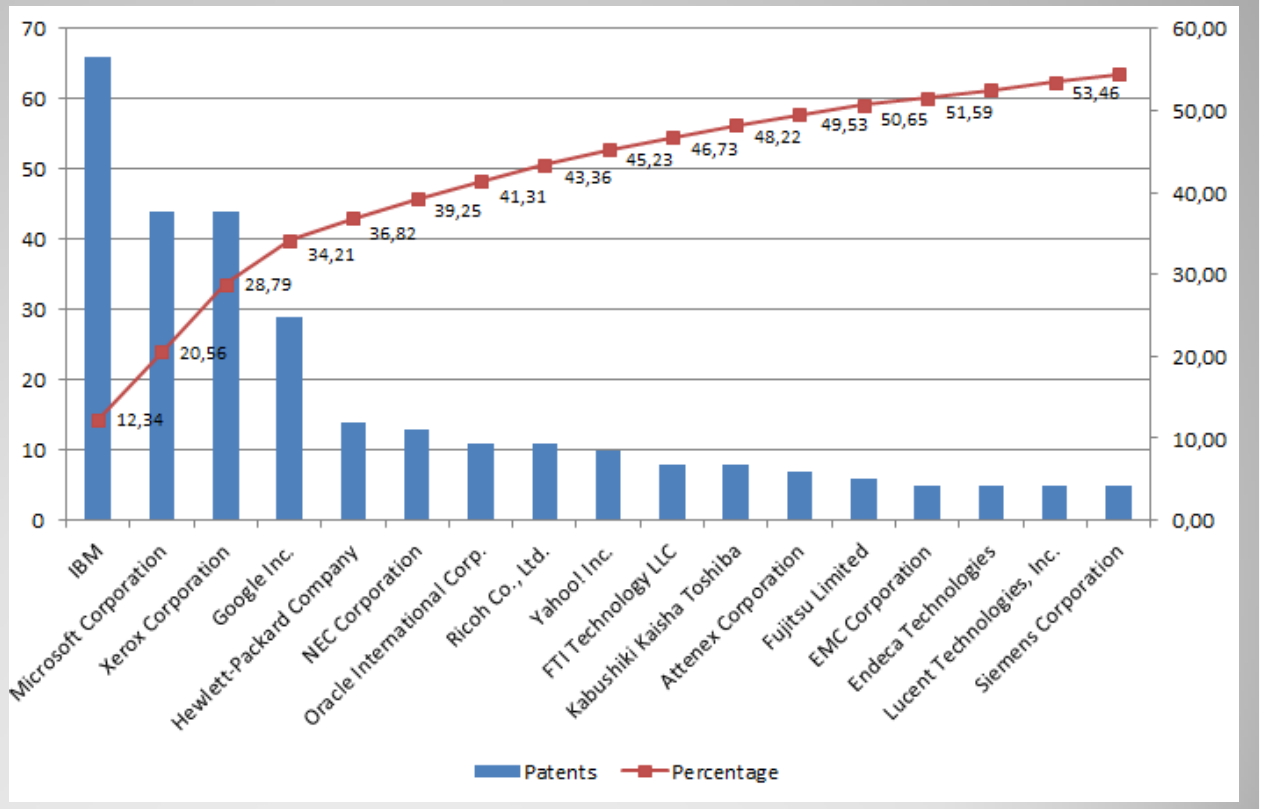
Distribution of patents by companies.

A few companies create most of the patents in this knowledge area.

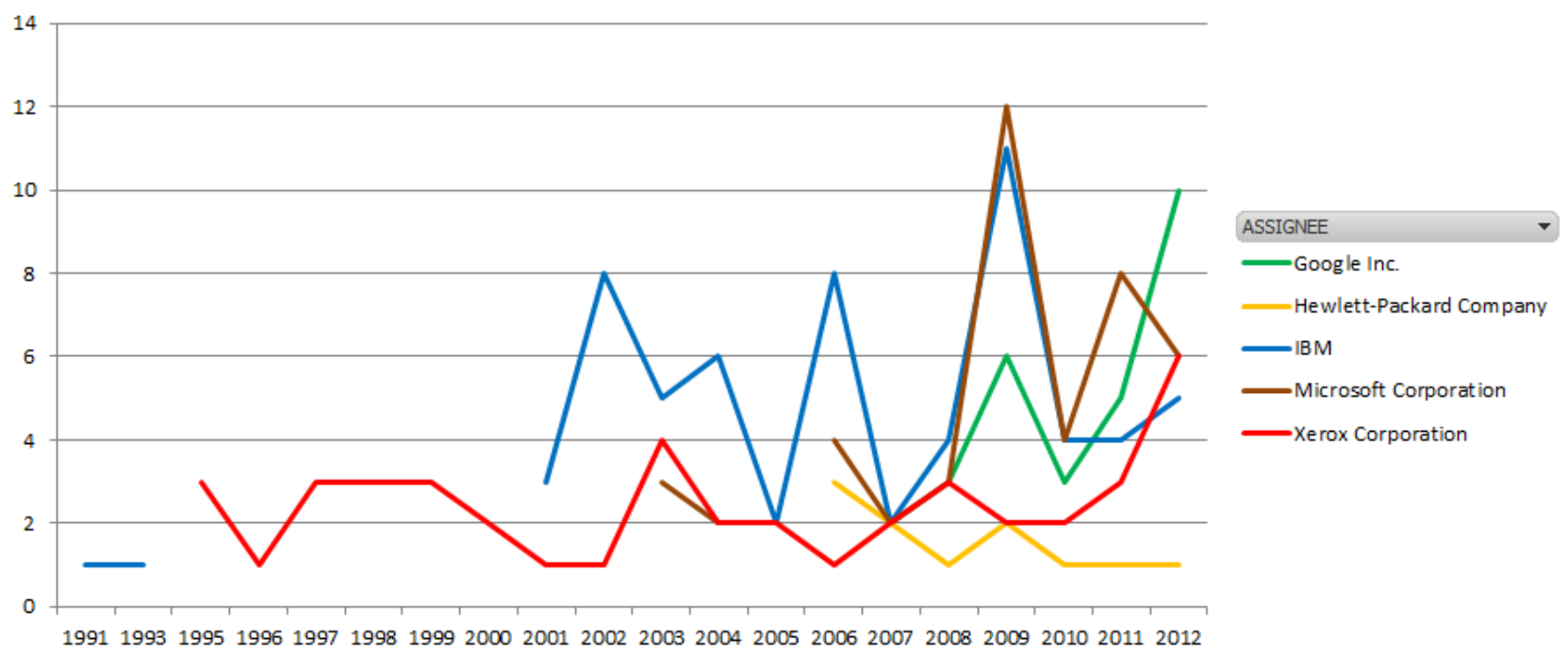
Similar to the classical distribution of academic articles in journals.

Sample Analysis
Which are the most "productive" companies?

ASSIGNEE	Total
IBM	66
Microsoft Corporation	44
Xerox Corporation	44
Google Inc.	29
Hewlett-Packard Company	14
NEC Corporation	13
Oracle International Corp.	11
Ricoh Co., Ltd.	11
Yahoo! Inc.	10
FTI Technology LLC	8
Kabushiki Kaisha Toshiba	8
Attenex Corporation	7
Fujitsu Limited	6
EMC Corporation	5
Endeca Technologies	5
Lucent Technologies, Inc.	5
Siemens Corporation	5

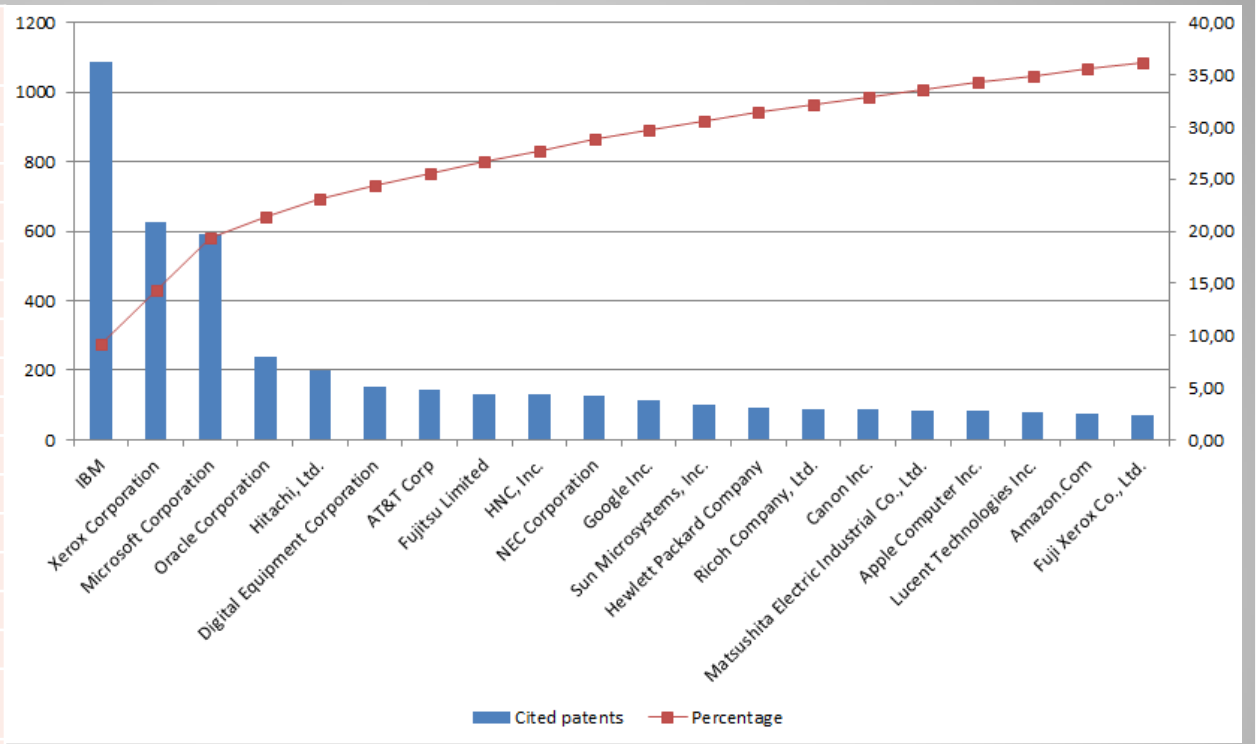


Sample Analysis
Evoluation of the most "productive" companies?



Sample Analysis of citations
Which are the most "influential" companies?

Company	Cited patents
IBM	1087
Xerox Corporation	625
Microsoft Corporation	593
Oracle Corporation	239
Hitachi, Ltd.	199
Digital Equipment Corp.	151
AT&T Corp	144
Fujitsu Limited	129
HNC, Inc.	129
NEC Corporation	126
Google Inc.	115
Sun Microsystems, Inc.	99
Hewlett Packard Company	92
Ricoh Company, Ltd.	89
Canon Inc.	86
Matsushita Electric Industrial Co., Ltd.	84
Apple Computer Inc.	83
Lucent Technologies Inc.	78
Amazon.Com	77
Fuji Xerox Co., Ltd.	72

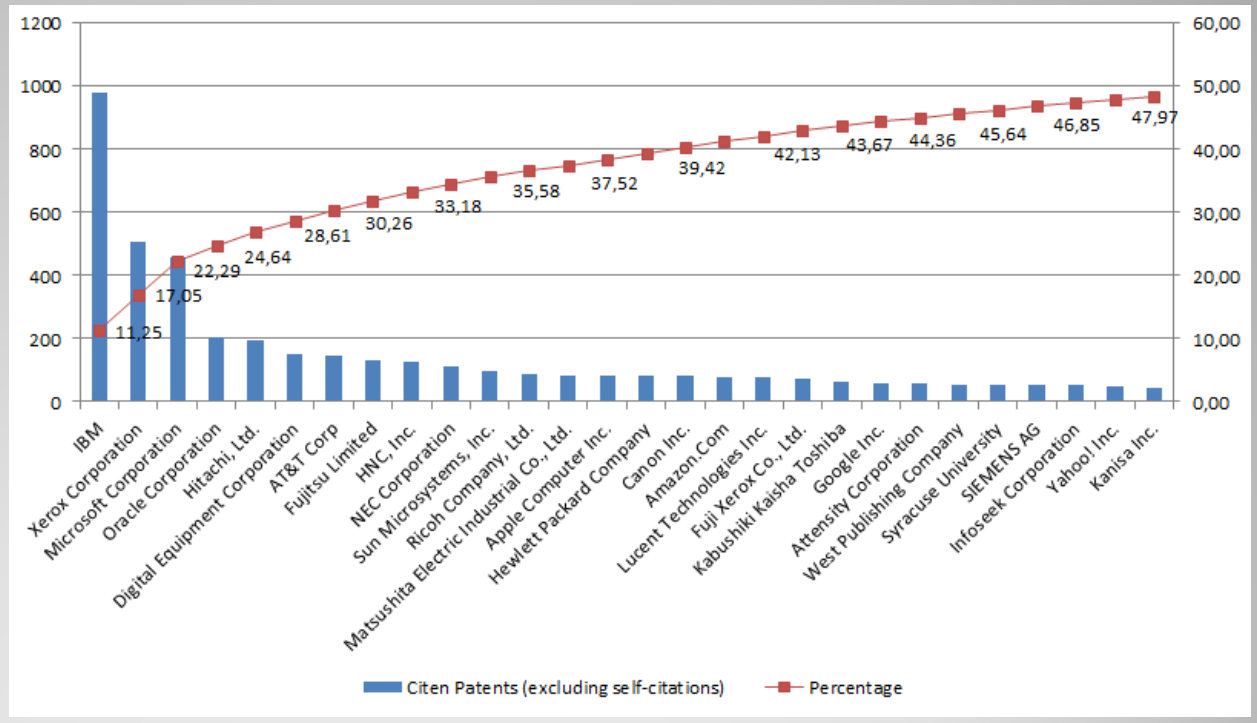


Impact based on citations received by their patents.

Identify companies with higher impact but lower production.

Sample Analysis of citations
Which are the most "influential" companies, excluding self-citation?

Assignee	Citen Patents
IBM	981
Xerox Corporation	505
Microsoft	457
Oracle	205
Hitachi, Ltd.	195
Digital Equipment Corp	151
AT&T Corp	144
Fujitsu Limited	129
HNC, Inc.	126
NEC Corporation	111
Sun Microsystems, Inc.	98
Ricoh Company, Ltd.	85
Matsushita Electric Industrial Co., Ltd.	84
Apple Computer Inc.	83
Hewlett Packard	83
Canon Inc.	82
Amazon.Com	77
Lucent Technologies	77
Fuji Xerox Co., Ltd.	71
Kabushiki Kaisha Toshiba	63



Impact of self-citation on data set does not seem to be relevant.

Ranking of companies is not affected by its removal.

Sample Analysis
Evolution of the most influential companies

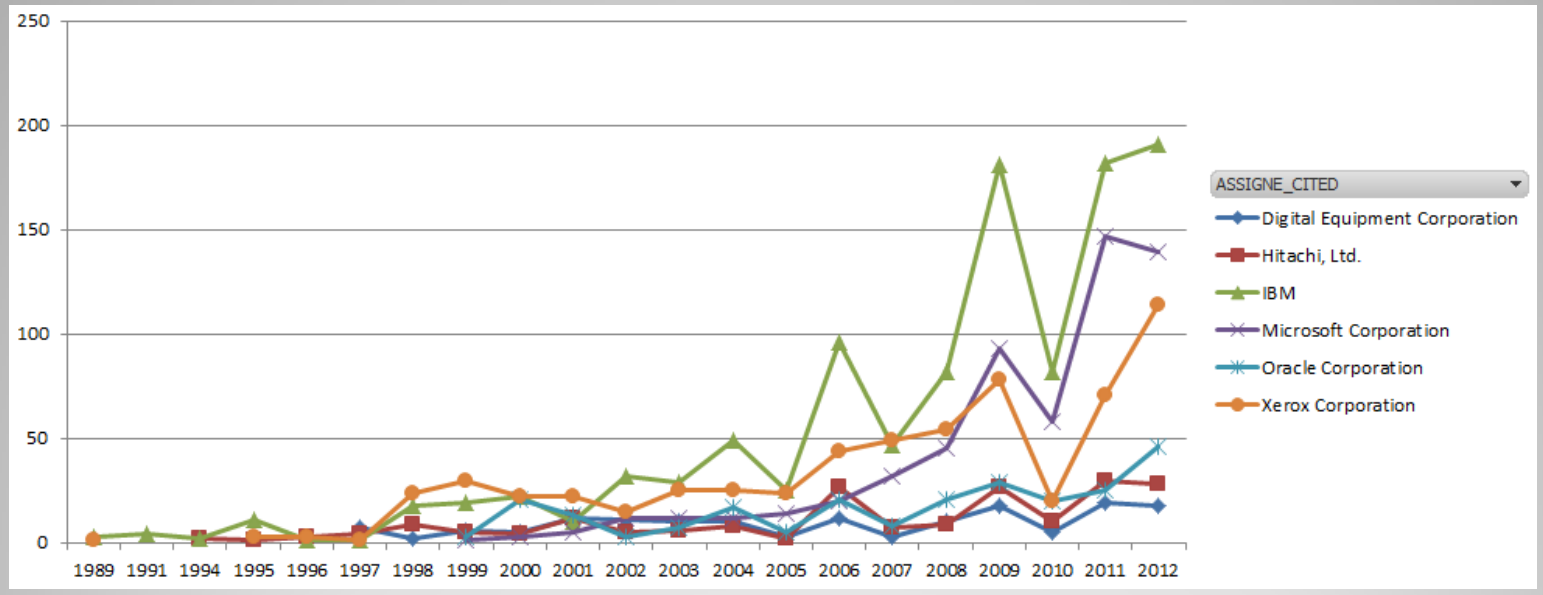


Figure includes self-citation, but self-citation does not have an impact on this figure.

With the exception of 2010, number of citation grows in the 2005-2012 period.

Sample Analysis
Bradford and the "core producers"

- Bradford's law is a classical bibliographical method initially proposed to identify the most important journals in a specific area.
- Bradford analysis is about "dispersion" of relevant literature in a collection of journals.
- It states that there is an "*inverse relationship between the number of articles published in a subject area and the number of journals in which the articles appear*".
- Bradford analysis identifies the "core" set of journals, based on the number of citations they receive from articles published in the area.
- Its objective was helping librarians decide to which journals the library should subscribe (better investment of budget for acquisitions).

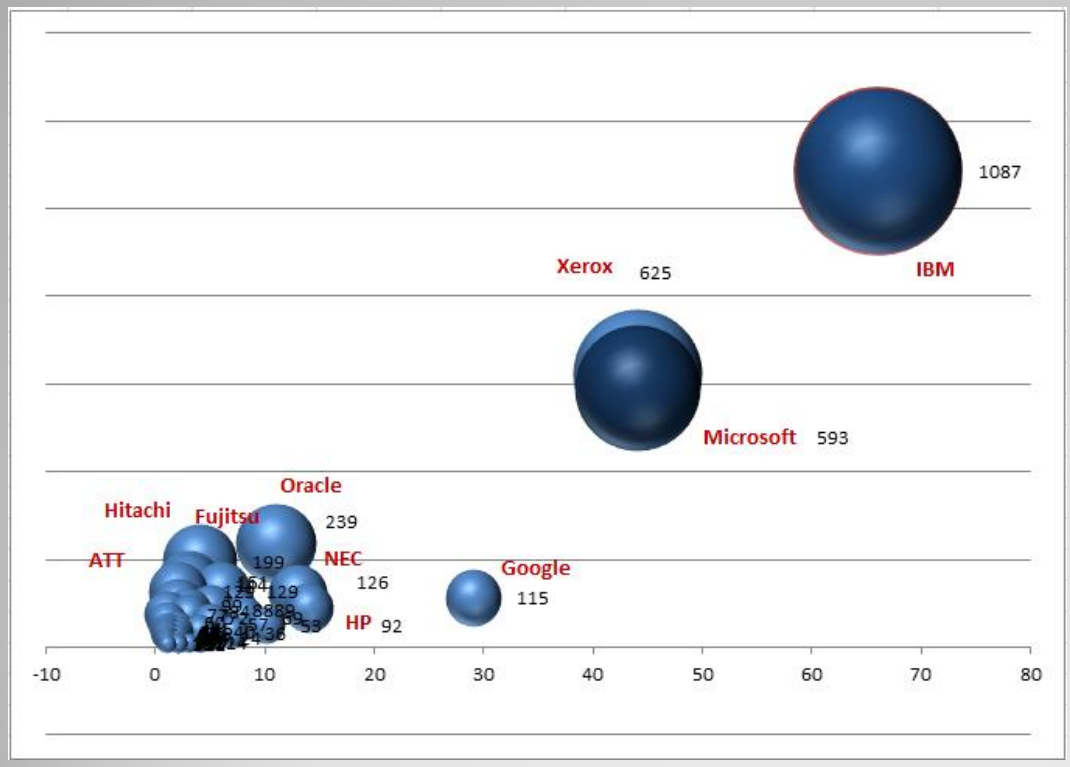
Sample Analysis
Bradford and the "core producers"

- Bradford law is also applied to asses authors, universities, etc.
- By applying Bradford analysis to patent citations, it is possible to identify the "core companies" generating contributions/inventions to a specific area.
- Companies are divided into three or more zones, each zone with the same number of citations.
- For a distribution in 4 zones, with around 2300 citations:
  - Core is made of 5 companies: IBM, Xerox, Microsoft, Oracle, Hitachi.
  - 2nd Zone includes 28 companies
  - 3rd Zone includes 185 companies
  - 4th Zone includes 1195 companies
  - Each zone increments number of assignees following the pattern $1:n:n^2:n^3\ldots$

Sample Analysis
Productivity and impact

- Bibliometric studies need to relate productivity (number of published items) with impact (citations received by the published items)
- One company may have a big number of patents with a small number of citations, or a small number of patents with a big number of citations…
- How can we put together these two variables?
- To deal with that, additional metrics have been provided by the bibliometric community: h-index, impact factor, g-index…
- A preliminary analysis of productivity and impact has been conducted for the sample dataset.

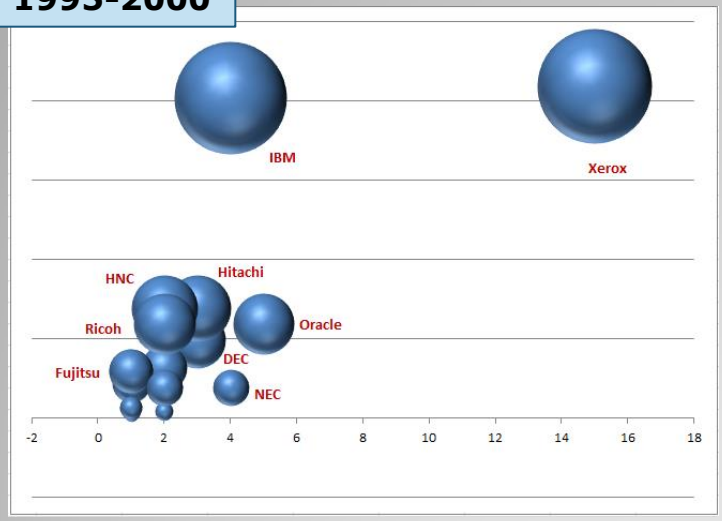
Sample Analysis
Evolution of the most influential companies



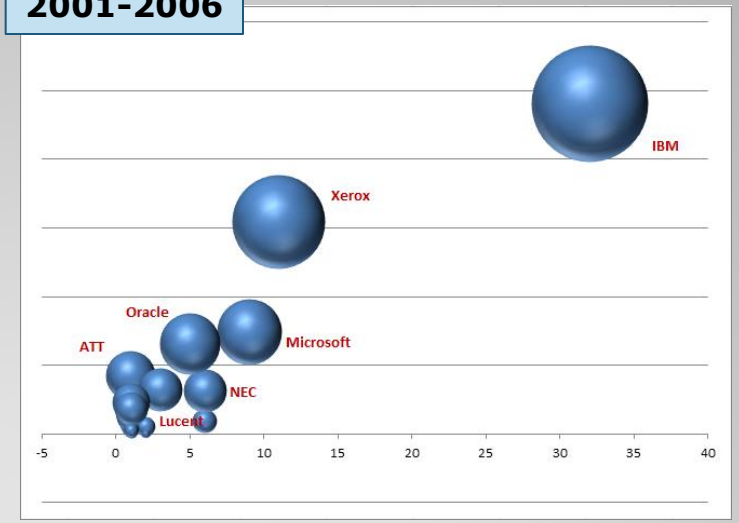
- This chart shows both productivity and impact.
- X-axe represents productivity (# patents) of the company in the period (1995-2012).
- Y-axe and the size of the bubbles represent the impact of the company in the period (received citations).
- It is possible to create this chart for different periods to analyze the evolution of companies in time.

Sample Analysis
Evolution of the most influential companies

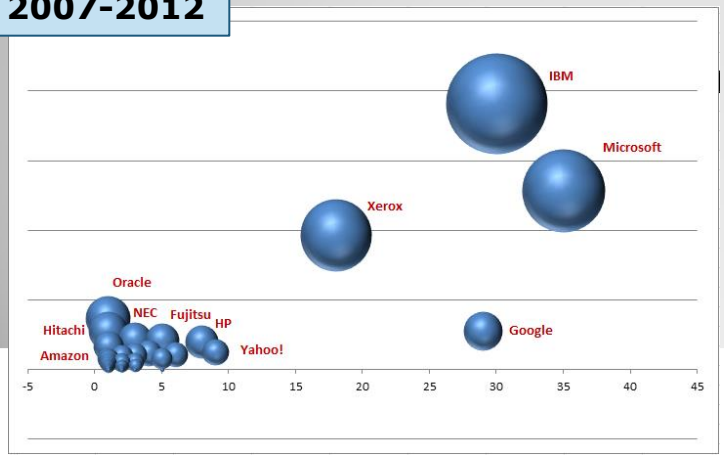
1995-2000



2001-2006

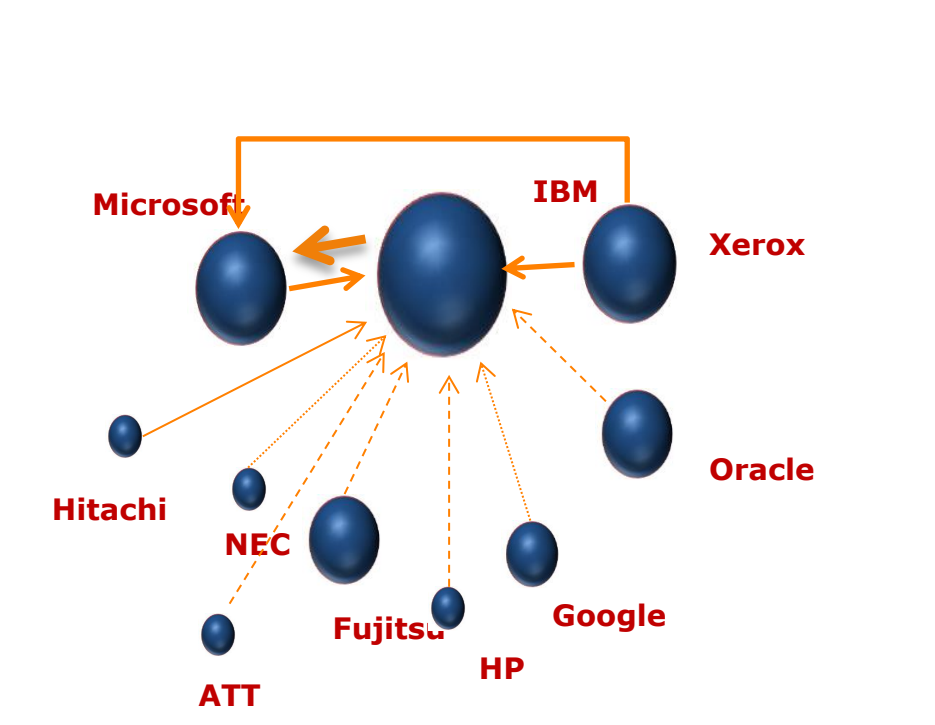


2007-2012



Evolution of companies' trends could be represented as vectors in a 2-dimensional space.
The vector shows the evolution of the company regarding production and impact.
This could provide a dynamic view of companies' innovation outputs.

Sample Analysis
Impact diagrams



- Another interesting output shows the impact that companies have on another companies.
- Classical "citation graphs".
- Arrow size demonstrate the impact of the relationship.
- To be generated for the companies with greater impact and for specific companies.
- Size of the arrows represent a weighted metric based on citations made, divided by total number of citations

Sample Analysis
Other bibliometric indicators

- The relationship between productivity and impact has been assessed with the h-index (Hirsch Index)
- "*It quantifies the cumulative impact and relevance of the scientific output of an individual*".
- Index is h if h of his N papers have at least h citations each and the other (N-h) papers have <= h citations each.
- H-Index takes into account both quantity of papers and the citations these papers have received.
- It gives a single number which measures the broad impact of an individual works, and allows authors to be compared according to their h-index.

Sample Analysis
Other bibliometric metrics

- For the sample data set, h-index obtained are the following:
  - 5 – IBM, XEROX
  - 4 – DEC, Infoseek
  - 3 – Microsoft, Fujitsu, Amazon, Canon, HNC, ATT
  - 2 - Intel, HP, Apple, Oracle, Hitachi, Yahoo!, Toshiba, Google, Accenture, Lexis-Nexis, Lucent, Lycos, MIT, SAP, Syracuse University, University of California…
- H-index is also used to calculate the "core" patents of an organization, those that had more impact on later research.
- H-index is dynamic and evolves with time, so it has to be monitored.

Sample Analysis
Other bibliometric metrics

- H-index is insensitive to the set of non-cited or lowly cited papers, and also to the set of highly cited papers.
- Is this an advantage or a limitation?
- Egghe proposed that insensitivity to lowly cited papers is right, but the index should be sensitive to highly cited papers.
- Egghe indicates that the number of citations received should be taken into considerations as a metric of the overall quality.
- G-Index is the "*unique, largest number such that the top g papers together receive g^2 or more citations, consequently g>=h.*"

Sample Analysis
Other bibliometric metrics

- For the sample data, h-index are the following:
  - 6 – IBM, XEROX
  - 5 – Microsoft, HNC, ATT
  - 4 – Fujitsu, Amazon, SAP
  - 3 – Oracle, Canon, Hitachi, Yahoo!, Intel, HP, Syracuse Univ.
  - 2 – DEC, Google, Lexis-Nexis, Toshiba, Accenture, Apple, MIT
  - 1 – Infoseek, Lycos, Lucent
- G-index is also used to calculate the "core" patents of an organization, those that had more impact on later research.

Sample Analysis
Other bibliometric metrics

- Immediacy index:
  - *Indicates the speed with which published items are incorporated into other references.*
  - *A high immediacy index indicates that the content is quickly noticed, highly valued and topical within the field of study.*
- Calculated as:
  - (Number of citations given to items in a year) /
  - (Number of items published in that year)

Sample Analysis
Other bibliometric metrics

- Impact Factor:
  - *Also developed to identify the most relevant journals in an area of study and facilitate journal selection using objective quantitative methods.*
  - *Sorting journals by impact factor enables de includion of many small but influential journals.*
  - *Annually calculad in JCR.*
  - *Applied to assess impact of authors, groups, academic departments and disciplines.*
- Calculated as:
  - (Citations to recent items during the year) /
  - (Number of recent items published)
- Recent mean "published in the last 2 or 5 years)

Sample Analysis
Other bibliometric metrics

- Impact Factor:
  - Impact factor includes self-citation, although initially is was used with the previous removal of self-citation.
  - The time period used for analysis is 2 or 5 years (target window), as different disciplines have different citation patterns, and 2 years may be a short time in most of the cases.
  - In the case of patents, a larger target window seems to be a better option, although the two indexes can be used together.

Sample Analysis
Other bibliometric metrics

- PEI (Publication Efficiency Index):
  - It assesses whether the impact of publications in a country in a given research field is compatible with its research effort.
  - PEI greater than 1 means that the impact of publication in this country is greater than the research effort made.
  - Analysis based on the ration of citations received per item published by a country compared to this ration for al the countries included in the analysis (per year).
  - The most productive countries are not necessarily those which obtain higher PEI values.
  - PEI = (TNCi / TNCt) / (TNPi / TNPt)
  - TNCi = total number of citations received by country
  - TNCt = total number of citations received by all countries
  - TNPi = total number of items published by country I
  - TNPt = total number of items published by all the countires.

Sample Analysis
Other bibliometric metrics

- Eigenfactor metrics:
  - Includes 2 metrics based on citation data to assess the influence of a journal in relation to other journals: Eigenfactor score and Article Influence score.
  - Are based on the idea that connections in scientific literature are made by citations, and remedy the biases associated with impact factor: the failure to take into account the differences in prestige between citing journals, and the difference sin citation patterns across disciplines.
  - Considers citations received by journals in the last 5 years and excludes self-citations.
  - It has been applied to other documents: thesis, books, newspapers…

Sample Analysis
Impact of journals on data set

- By applying previous techniques, the most influencial journals on patents have been identified (by citations):
- Journals and conferences in the core were the following one:
  - ACM Annual Conf. On R&D in Inf.Retrieval - SIGIR
  - ACM Transactions on Information Systems
  - TREC
  - ARPA Workshop on Human Language Technology
  - Communications of the ACM
  - Int.Conf.on Machine Learning
  - ACM SIGCHI Computer-Human Interaction
  - Information Processing & Management
  - National Online Meeting
  - Computer Assisted Inf. Searching on Internet RIAO

Conclusions

- Patent citation analysis may be used to identify groups focused on innovation, both in companies and in universities
- In the sample data, big figures hide the effort and outputs from universities, but they can be treated as a subset.
- The analysis is valid to identify academic journals and other publications that had an impact on the development of the innovations.
- Perhaps this analysis may give some answers to the classical problem: which is the actual, practical value of academic research? How can we measure that?
- The assessments of academic institutions is today a key topic regarding R&D policies, and it should not focus exclusively on their output in academic journals.

Conclusions. Next Steps

- Identify a wider set of patents for the Text Mining area, to include patents related to information extraction and automatic text summarization.
- Complete the analysis with additional bibliometric indicators.
- Analyze and assess how the classical bibliometric indicators reflect the evolution of each organization on the "productivity and impact" scenario.
- Repeat the analysis for additional areas and domains.
- Identify additional sources of data – in addition to Delphion and the US Patents collection – to validate the results.

Thanks!!

- Questions:

Ricardo Eito-Brun

reito@bib.uc3m.es

Note: Definitions for bibliometric indicators are taken from: ANDRÉS, Ana. Measuring Academic Research. Chandos Publishing, 2009