AKBULUT, Müge. The Analysis of the Impact of Citation Classics and Relevance Rankings Using Pennant Diagrams (in Turkish). Unpublished Master's thesis, Hacettepe University, Ankara, 2016. (Available at: http://www.mugeakbulut.com/yayinlar/Muge_Akbulut_YL_Tez.pdf)

Extended Abstract

Introduction

Citation indexes are important tools to measure the impact of scholarly publications. Yet, they tend to work rather poorly for older publications and fail to reveal the true impact of such contributions. This is especially the case for so called "citation classics" that were published before citation indexes came into being.

We address this problem using a seminal work on probabilistic indexing by Maron and Kuhns (1960) as a case study (hereafter M&K). M&K's paper has replaced the "two-valued thinking about information retrieval with probabilistic notions" (Maron, 2008, p. 971), thereby paving the way, almost 60 years ago, for Google-like systems that we take for granted today.

M&K's paper came out before the citation indexes were launched and therefore did not gather as many citations as it deserved even though it is still a "citation classic" by definition with over 400 citations. However, raw citation data does not truly reflect the impact of M&K's paper on information retrieval models developed later.

This thesis investigates both direct and indirect influence of M&K's paper using pennant diagrams based on author co-citation analysis (ACA), information retrieval (IR) and relevance theory (RT). ACA was first introduced by White and Griffth (1981) to measure the intellectual structure of a discipline. More recently, White (2007a, 2007b) combined ACA with IR and RT to study the impact of a core work, author or paper (such as that of M&K). He used the classical *tf***idf* (term frequency*inverse document frequency) formula to create pennant diagrams.

Originally, *tf* and *idf* values are used to represent term relevance within a document and a collection, respectively, on the basis of term frequency within a document and all the documents in the corpus. White used *tf***idf* formula "to mimic a relevance theoretic model of cognition on the user side" (White, 2007a). More specifically (using M&K' paper as an example), *tf* values represent "cognitive effects" of other papers in the context of M&K's paper, and *idf* values represent the "processing effort" of the user (i.e., ease of access). White used *tf***idf* values separately to plot pennant diagrams in which the *x* and *y* axes represent *tf* (cognitive impact) and *idf* (ease of access) values, respectively (Tonta & Özkan Çelik, 2013, p. 39). Thus, one can easily discern the relationship between, say, M&K's paper and the cited papers in terms of cognitive impact versus ease of access on a scatterplot and trace the intellectual structures of disciplines (White, 2015).

As similar papers tend to cluster, pennant diagrams also reveal the interdisciplinary relationships between M&K's paper and the (co-)cited papers. Moreover, as pennant diagrams take into account both direct and indirect citations to M&K's paper, relevance rankings based on pennant diagrams should produce better results than the "Related records" feature currently used in citation indexes, which is based on the similarity of reference lists of papers.

We conjectured that using pennant diagrams (a) the impact of M&K's paper can be studied more thoroughly; (b) the interdisciplinary relations that are unobservable with traditional citation analysis can be visualized temporally; and (c) relevance rankings based on pennant diagrams will produce more relevant "related records" for M&K's paper than the current ones. Finding answers to these questions will shed some light on the true impact of M&K's paper and relevance rankings based on similarity of reference lists.

Data and Methods

To address these questions, we identified through Web of Science all the works that cited or co-cited M&K's paper between 1960 and 2015 and downloaded them including their references (4,176 unique works). We wrote several macros to clean, merge, cluster, count and visualize data. We computed the *tf* and *idf* values as follows:

tf = (1 + log (*tf*)) (represents *x* axis, "cognitive impact")

idf = (log (*N/df*) (represents *y* axis, "ease of access").

*tf*idf* = *relevance* = (1 + log (*tf*)) * (log (*N/df*)

where *tf* and *idf* values are based on the number of co-citations and total number of citations, respectively. These values were later used to plot interactive pennant diagrams and visualize the data using CiteSpace (Chen, 2006). We wrote a macro to match similar records and used the similarity threshold of 80%. The output was also visually examined prior to merging citations. We created the static pennant diagrams and plotted the interactive ones using *Time Chart* available through Google Developers source code platform along with the Google API. We used a macro to count the number of total (co)-citations for each of the 4,176 records based on the similarity of their reference lists and computed the *tf*idf* weightings accordingly to come up with relevance rankings. We selected 90 records using convenience sampling techniques to create interactive pennant diagrams for further analysis.

Results and Discussion

The pennant diagram in Fig. 1 shows the impact of M&K's paper (labeled "core paper" on the right-hand side of the diagram). Each dot represents a paper. Clusters of papers are labeled according to their topical similarities to the core paper and to each other. The ones closer to the upper right-hand quadrant of *x* axis have been co-cited with the core paper more often (e.g., Boolean, vector space and probabilistic IR models) and had more cognitive impact. As the scatterplot shows, their numbers are relatively few.

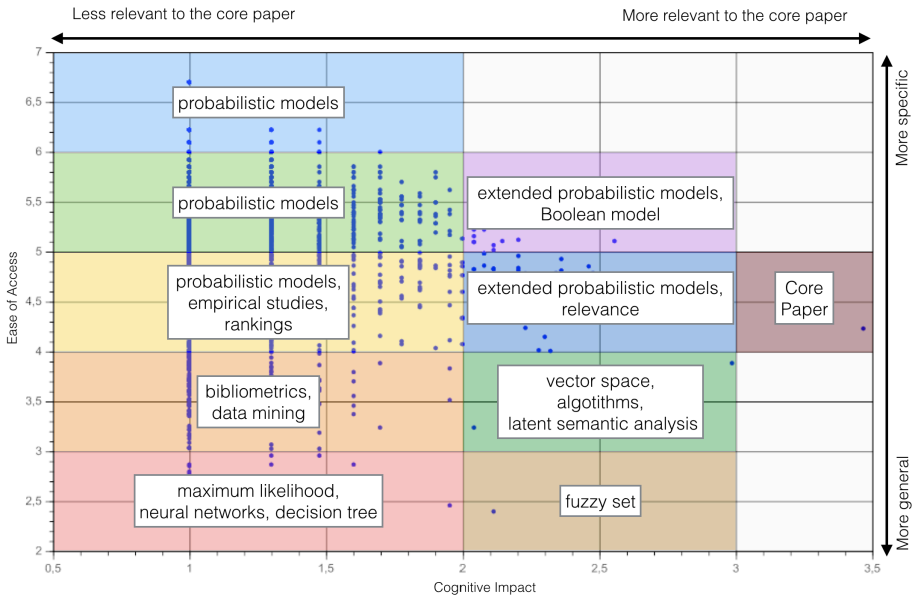

FIG. 1. Distribution of studies by topics

As we go farther away from the core paper to the left, the number of co-citations drops while the number of total citations increases. Papers in the upper left-hand quadrant of the *y* axis are easier to get access to because of their higher total number of citations but they have presumably had much less

cognitive impact on the topic of core paper. More general papers are generally close to the bottom half of the *y* axis.

Fig. 2 is a screenshot of the interactive pennant diagram showing the distribution of citations by years between the publication year of the core paper (1960) and 2015, and can be "played" online at (https://goo.gl/Wjr1YD). By running it, one can follow the relationship between the core paper and others in terms of cognitive impact and ease of access temporally. Fig. 2 also shows the topical relevance of those papers to the core paper. For instance, 81% of the papers with "probabilistic" in their titles appeared in the upper part of the diagram labeled as "A". These are directly related with the core paper's topic. The relevance of more general works in sectors B and C to the core paper is less obvious but they are relevant nonetheless (e.g., Shannon and Weaver's book entitled *The Mathematical Theory of Communication* appears in section C).

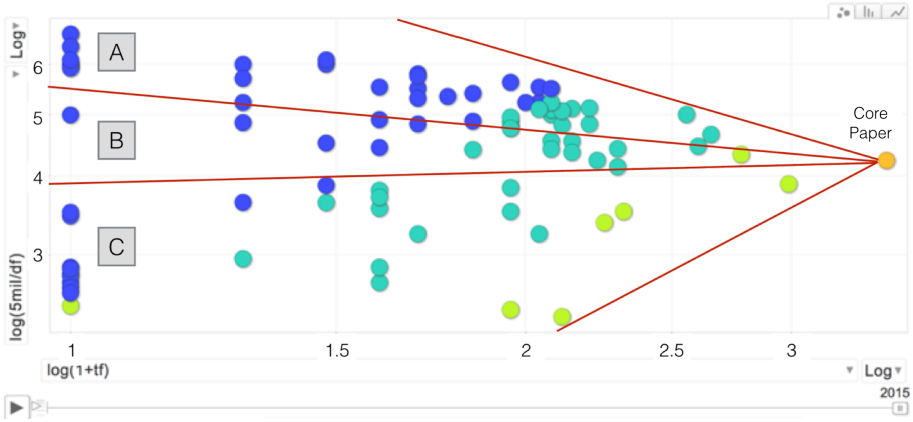

FIG. 2. Screenshot of interactive pennant diagram

Pennant diagrams show the influence of a given work, model or paradigm on other works in a specific field or on other fields in general. They can also help researchers track the relevant literature on a topic more easily along with its emergence and evolution. We obtained promising results for relevance rankings as well. M&K's paper had only two citations: one to Shannon and Weaver's book, and the other to Yule's classic paper "on measuring association between attributes" (1912). None of the "Related records" suggested by citation indexes on the basis of the similarity of reference lists was relevant to M&K's paper whereas all the records suggested by the pennant diagram were highly relevant. As far as we know, this is the first study comparing the performance of current relevance rankings with that of pennant diagrams. The lists of suggested works by the pennant diagram and citation indexes are available for comparison in Appendix 3 (pp. 85-87) of the thesis (http://www.mugeakbulut.com/yayinlar/Muge_Akbulut_YL_Tez.pdf). The one produced by pennant diagram can be readily used as a "reading list" for courses on IR models and IR history.

Our study is based on but a single case. Yet, pennant diagrams combining bibliometrics, IR and relevance theory as suggested by White seem to offer a promising avenue to study the impact of citation classics and trace the underlying intellectual structure of the relationships between influential works. Alternative relevance rankings based on pennant diagrams can be offered to users and integrated into the current citation databases and recommender systems.

References

Chen, C. (2006). CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. *JASIST,* 57(3), 359–377.

Maron, M.E. (2008). An historical note on the origins of probabilistic indexing. *Information Processing and Management*, 44, 971-972.

Maron, M.E. & Kuhns, J.L. (1960). On relevance, probabilistic indexing and information retrieval. *Journal of the ACM,* 7, 216-244.

Tonta, Y. & Özkan Çelik, A.E. (2013). Cahit Arf: Exploring His Scientific Influence Using Social Network Analysis, Author Co-citation Maps and Single Publication h Index. *Journal of Scientometric Research*, 2(1): 37-51.

White, H.D. (2007a). Combining bibliometrics, information retrieval, and relevance theory. Part 1: First examples of a synthesis. *JASIST*, 58, 536-559.

White, H.D. (2007b). Combining bibliometrics, information retrieval, and relevance theory. Part 2: Some implications for information science. *JASIST*, 58, 583-605.

White, H.D. (2015). Co-cited author retrieval and relevance theory: examples from the humanities. *Scientometrics*, 102(3), 2275-2299.

White, H.D. & Griffith, B.C. (1981). Author co-citation: A literature measure of intellectual structure. *JASIST*, 32, 163-171.