Enhancement of Text Representations Using Related Document Titles‡

G. Salton*
Y. Zhang†
TR 86-728
January 1986

Department of Computer Science
Cornell University
Ithaca, NY 14853

‡ This study was supported in part by the National Science Foundation grant IST 83-16166.
* Department of Computer Science, Cornell University, Ithaca, NY 14853.
† Institute of Computer Technology, China Academy of Railway Sciences, Beijing, China.

Enhancement of Text Representations Using Related Document Titles

G. Salton* and Y. Zhang**

Abstract

Various attempts have been made over the years to construct enhanced document representations by using thesauruses of related terms, term association maps, or knowledge frameworks that can be used to extract appropriate terms and concepts. None of the proposed methods for the improvement of document representation has proved to be generally useful when applied to a variety of different retrieval environments. Some recent work by Kwok suggests that document indexing may be enhanced by using title words taken from bibliographically related items. An evaluation of the process shows that many useful content words can be extracted from related document titles, as well as many terms of doubtful value. Overall, the procedure is not sufficiently reliable to warrant incorporation into operational automatic retrieval systems.

---

*Department of Computer Science, Cornell University, Ithaca, NY 14853.

**Institute of Computer Technology, China Academy of Railway Sciences, Beijing, China.

This study was supported in part by the National Science Foundation under grant IST 83-16166.

1. Term and Document Relations

Most existing methods for the automatic content analysis of written texts are based in part on the extraction of certain words contained in the original document texts. While many words appearing in ordinary text are in fact useful for content representation, it is often believed that the use of text words does not provide a complete description of text meaning. For this reason, various additional content analysis tools have been introduced in the hope of obtaining more complete text representations. Among these tools are thesauruses that contain groupings of related words [1,2], automatically constructed term association maps based on co-occurrences of words in the texts of documents [3,4], and knowledge frameworks representing the facts and relationships that characterize particular subject areas. [5-7]

Various methodologies have been suggested to help in the construction of the content analysis tools, including for example probabilistic theories of information processing that account for the use of term relationships and associations [8-10], methods that include syntactic considerations for the construction of term phrases [11-13], and finally interactive procedures in which individual users may suggest term relationships of importance in their application based on a dialogue between user and system conducted from a user terminal. [14-16]

Two main problems arise when term associations are proposed for text identification and processing:

a) No theory exists which would help in distinguishing valuable term associations from less valuable ones, and no obvious help is available to aid in the construction of useful thesauruses, association maps,

DOCKET ALARM

Explore Litigation Insights

Docket Alarm provides insights to develop a more informed litigation strategy and the peace of mind of knowing you're on top of things.

Real-Time Litigation Alerts

Keep your litigation team up-to-date with **real-time alerts** and advanced team management tools built for the enterprise, all while greatly reducing PACER spend.

Our comprehensive service means we can handle Federal, State, and Administrative courts across the country.

Advanced Docket Research

With over 230 million records, Docket Alarm's cloud-native docket research platform finds what other services can't. Coverage includes Federal, State, plus PTAB, TTAB, ITC and NLRB decisions, all in one place.

Identify arguments that have been successful in the past with full text, pinpoint searching. Link to case law cited within any court document via Fastcase.

Analytics At Your Fingertips

Learn what happened the last time a particular judge, opposing counsel or company faced cases similar to yours.

Advanced out-of-the-box PTAB and TTAB analytics are always at your fingertips.

API

Docket Alarm offers a powerful API (application programming interface) to developers that want to integrate case filings into their apps.

LAW FIRMS
Build custom dashboards for your attorneys and clients with live data direct from the court.

Automate many repetitive legal tasks like conflict checks, document management, and marketing.

FINANCIAL INSTITUTIONS
Litigation and bankruptcy checks for companies and debtors.

E-DISCOVERY AND LEGAL VENDORS
Sync your system to PACER to automate legal marketing.