

Fortschritt-Berichte VDI

VDI

Reihe 10

Informatik/
Kommunikation

Nr. 863

Herwig Unger,
Mario M. Kubek

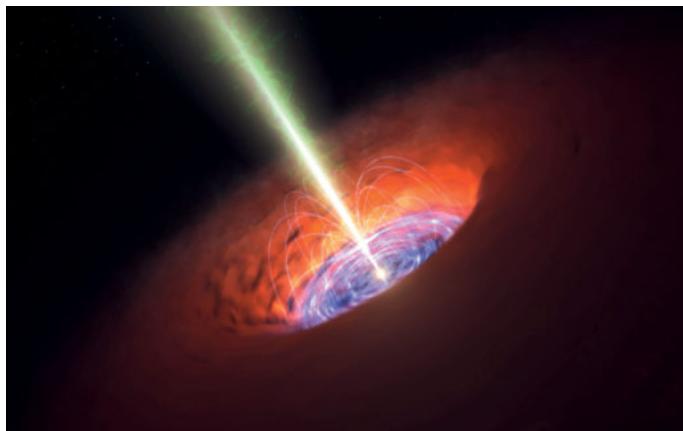
Theory and Application of Text-representing Centroids



FernUniversität in Hagen
Schriften zur Informations-
und Kommunikationstechnik

Theory and Application of Text-representing Centroids

Herwig Unger and Mario M. Kubek



Email: kn.wissenschaftler@fernuni-hagen.de
Website: <https://www.fernuni-hagen.de/kn/>

Fortschritt-Berichte VDI

Reihe 10

Informatik/
Kommunikation

Herwig Unger,
Mario M. Kubek

Nr. 863

Theory and Application
of Text-representing
Centroids



FernUniversität in Hagen
Schriften zur Informations-
und Kommunikationstechnik

Herwig Unger, Mario M. Kubek

Theory and Application of Text-representing Centroids

Fortschr.-Ber. VDI Reihe 10 Nr. 863. Düsseldorf: VDI Verlag 2019.

152 Seiten, 49 Bilder, 10 Tabellen.

ISBN 978-3-18-386310-5, ISSN 0178-9627,

€ 57,00/VDI-Mitgliederpreis € 51,30.

Keywords: Text Processing – Text Centroid – Co-occurrence Graph – Spreading Activation – Text Categorisation – Librarian of the Web – P2P-system – Decentralised Search – WebEngine – Web Search Engine

Centroid terms are single, descriptive words that semantically and topically characterise text documents and thus can act as their very compact representation in automated text processing tasks that strongly rely on the semantic similarity of texts. Algorithms to classify and cluster them make use of this information. In this book, the novel, brain- and physicsinspired concept of centroid terms is introduced and deeply discussed. Furthermore, their unique properties and practical usage in major natural language processing and text mining tasks are covered. In this regard, a new graph-based method for their fast calculation is presented as well. In contrast to methods relying on the bag-of-words model, the derived centroid distance measure can uncover a topical relationship between texts even when their wording differs. As centroid terms can also represent short texts, the presented first fully integrated, P2P-based web search engine, called "WebEngine", therefore makes heavy use of centroid terms when interpreting queries and forwarding them to peers with matching documents, represented by their own centroid terms.

Bibliographische Information der Deutschen Bibliothek

Die Deutsche Bibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliographie; detaillierte bibliographische Daten sind im Internet unter www.dnb.de abrufbar.

Bibliographic information published by the Deutsche Bibliothek

(German National Library)

The Deutsche Bibliothek lists this publication in the Deutsche Nationalbibliographie (German National Bibliography); detailed bibliographic data is available via Internet at www.dnb.de.

Schriften zur Informations- und Kommunikationstechnik

Herausgeber:

Wolfgang A. Halang, ehemaliger Lehrstuhl für Informationstechnik

Herwig Unger, Lehrstuhl für Kommunikationstechnik

FernUniversität in Hagen

© VDI Verlag GmbH · Düsseldorf 2019

Alle Rechte, auch das des auszugsweisen Nachdruckes, der auszugsweisen oder vollständigen Wiedergabe (Fotokopie, Mikrokopie), der Speicherung in Datenverarbeitungsanlagen, im Internet und das der Übersetzung, vorbehalten.

Als Manuskript gedruckt. Printed in Germany.

ISSN 0178-9627

ISBN 978-3-18-386310-5

Contents

| | |
|--|-----|
| Is a ‘Librarian of the Web’ really needed? | |
| H. Unger | 1 |
| Centroid Terms as Text Representatives | |
| M. M. Kubek and H. Unger | 7 |
| Spreading Activation: A Fast Calculation Method for Text Centroids | |
| M. M. Kubek, T. Böhme and H. Unger | 27 |
| Empiric Experiments with Text-representing Centroids | |
| M. M. Kubek, T. Böhme and H. Unger | 39 |
| Towards a Librarian of the Web | |
| M. M. Kubek and Herwig Unger | 55 |
| A Concept Supporting a Resilient, Fault-tolerant and Decentralised Search | |
| H. Unger and M. M. Kubek | 79 |
| An Associative Ring Memory to Support Decentralised Search | |
| H. Unger and M. M. Kubek | 91 |
| The WebEngine – A Fully Integrated, Decentralised Web Search Engine | |
| M. M. Kubek and H. Unger | 107 |
| On Evolving Text Centroids | |
| H. Unger and M. M. Kubek | 121 |
| Addendum | 131 |
| Authors | 140 |

