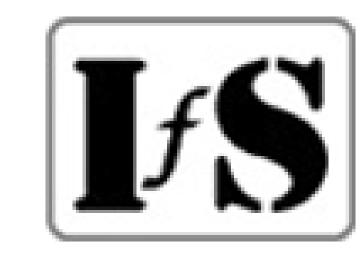


Masterstudium: Information Engineering and Knowledge Management

Automatische Zusammenfassung von Textclustern

Julius Penaranda



Technische Universität Wien Institut für Softwaretechnik und interaktive Systeme Arbeitsbereich: Information und Software Engineering Betreuer: Ao. Univ. Prof. Dipl.-Ing. Dr. techn. Andreas Rauber

Abstract

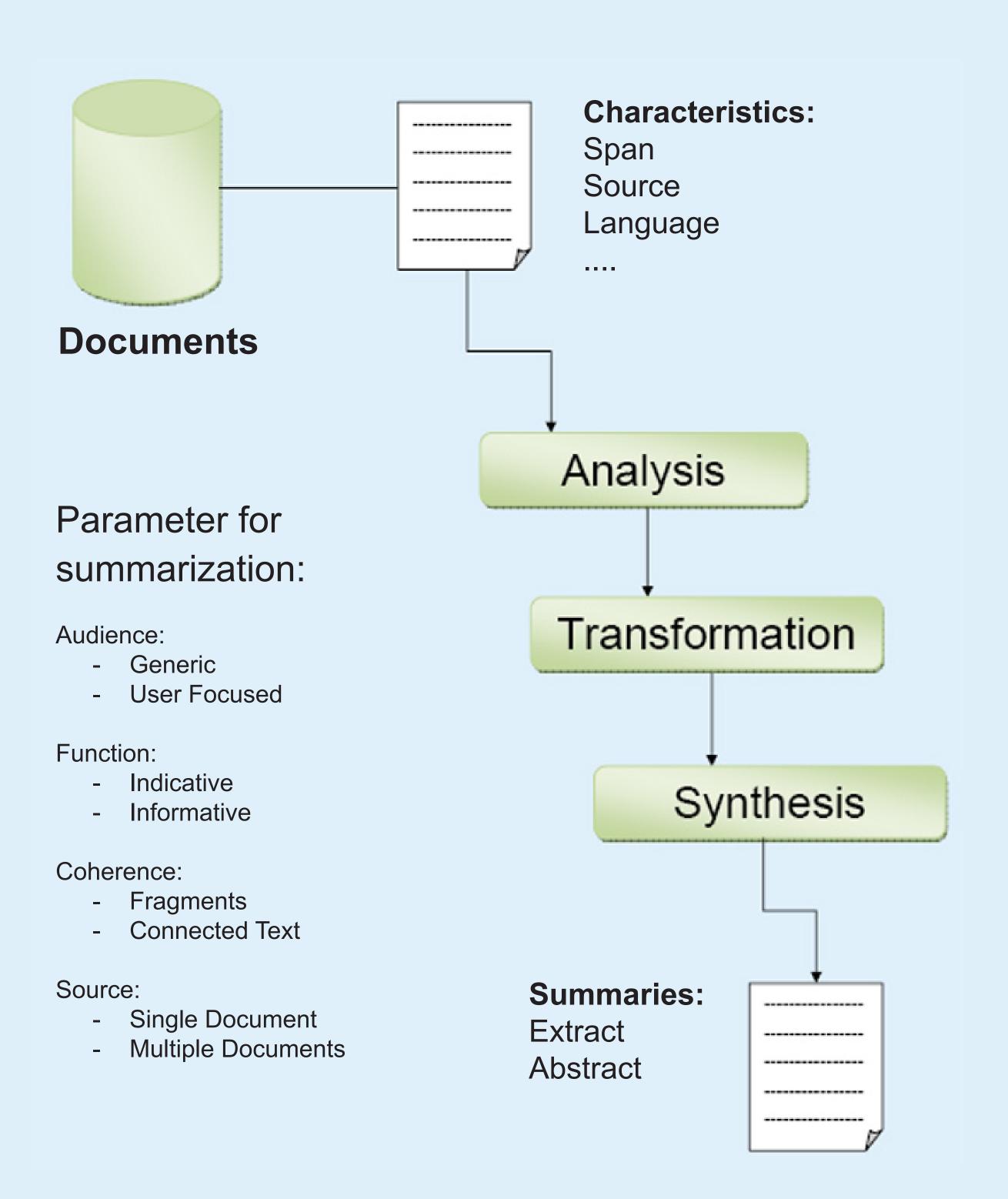
The internet offers a constantly growing amount of information to almost any desired topic, however, with the setback that it is time-consuming. The first search engines emerged so as to alleviate the problem of searching for relevant information. Nevertheless, it seems as though listings of the search results of relevant sites are no longer suitable. Thus, additional methods such as extraction algorithms for automatic summaries have been established to help reduce several texts on a topic to the essential contents. Consequently, the main points of the texts are rapidly acquired and clearer to the reader.

Definitions

Extract: a summary consisting entirely of material copied from the input

Abstract: a compressed and reformulated version of the contents of some portions of the text

Architecture



Implementation of various Extraction methods

Single-document Summarization

- Term Frequency method Keyphrase method
- Location method
- Title method

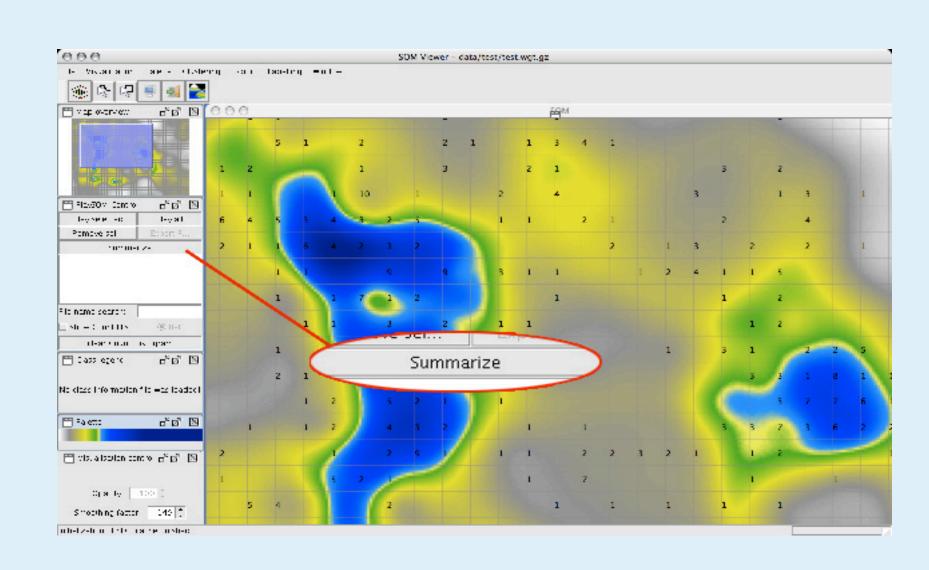
Multi-document Summarization

- Redundancy-based Algorithm:
- finds similarities and differences among documents

Summarization System

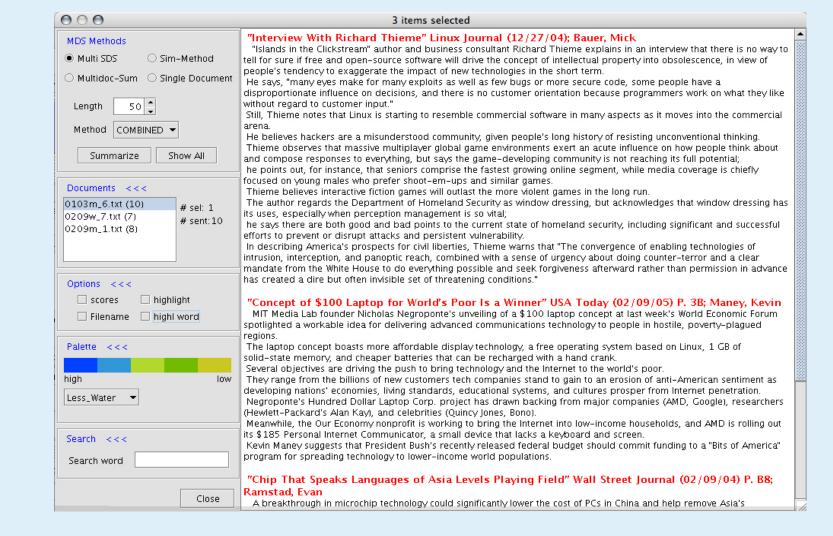
SOMToolBox

- extended by Summarization Tool
- Representation of data on a two-dimensional map that preserves topological information
- Exploration of collection of data
- Clustering of documents on a topic



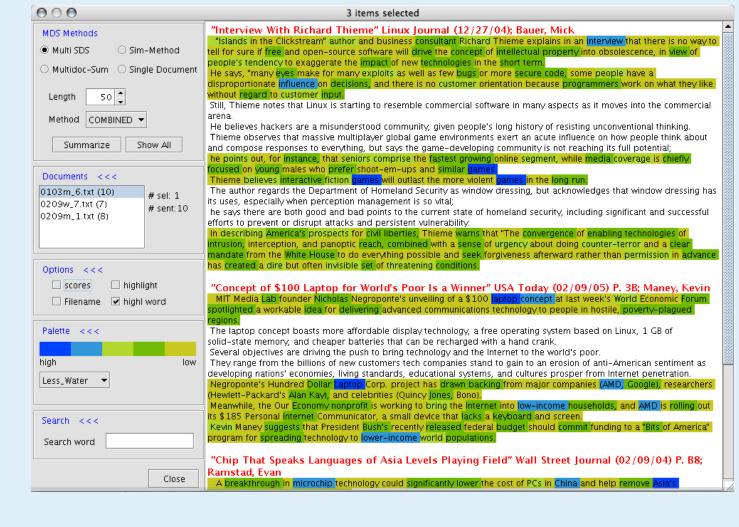
Summarization Tool

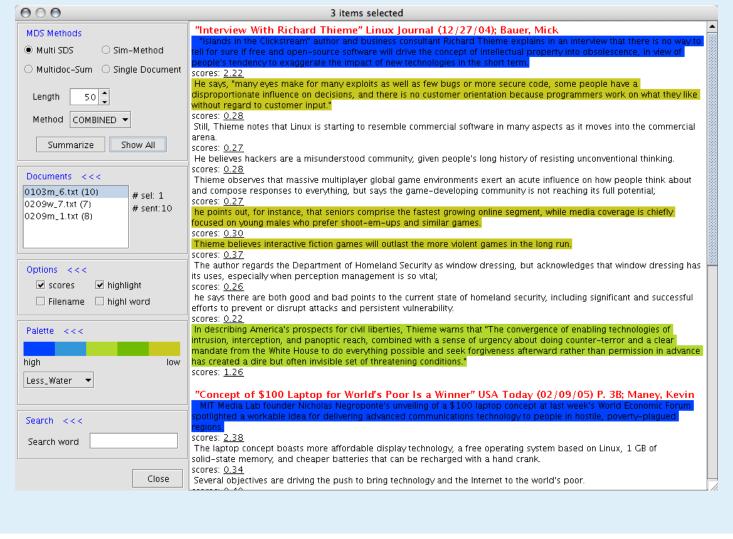
- Generation of single- and multidocument summarization - implemented in Java



Additional Features:

- highlights words in relevant sentences according to scores
- highlights relevant sentences according to scores
- displays scores of relevant sentences
- Search for words





000	3 items selected				
MDS Methods	"Interview With Richard Thieme" Linux Journal (12/27/04); Bauer, Mick				
Multi SDS Sim-Method	"Islands in the Clickstream" author and business consultant Richard Thieme explains in an interview that there is no wa tell for sure if free and open-source software will drive the concept of intellectual property into obsolescence, in view of people's tendency to exaggerate the impact of new technologies in the short term.				
○ Multidoc-Sum ○ Single Document	He says, "many eyes make for many exploits as well as few bugs or more secure code, some people have a disproportionate influence on decisions, and there is no customer orientation because programmers work on what they l				
Length 50 🗘	without regard to customer <mark>in</mark> put." Still, Thieme notes that L <mark>in</mark> ux is start <mark>in</mark> g to resemble commercial software <mark>in</mark> many aspects as it moves <mark>in</mark> to the commerc				
Method COMBINED ▼	arena. He believes hackers are a misunderstood community, given people's long history of resist <mark>in</mark> g unconventional th <mark>inkin</mark> g.				
Summarize Show All	Thieme observes that massive multiplayer global game environments exert an acute <mark>in</mark> fluence on how people th <mark>in</mark> k abou and compose responses to everythi <mark>n</mark> g, but says the game-develop <mark>in</mark> g community is not reach <mark>in</mark> g its full potential; he po <mark>in</mark> ts out, for <mark>in</mark> stance, that seniors comprise the fastest grow <mark>in</mark> g onl <mark>in</mark> e segment, while media coverage is chiefly				
Documents <<<	focused on young males who prefer shoot-em-ups and similar games. Thieme believes <mark>in</mark> teractive fiction games will outlast the more violent games <mark>in</mark> the long run.				
0103m_6.txt (10) # sel: 1	The author regards the Department of Homeland Security as w <mark>in</mark> dow dress <mark>in</mark> g, but acknowledges that w <mark>in</mark> dow dress <mark>in</mark> g				
0209w_7.txt (7) # sent:10 0209m_1.txt (8)	its uses, especially when perception management is so vital; he says there are both good and bad po <mark>in</mark> ts to the current state of homeland security, <mark>in</mark> clud <mark>in</mark> g significant and successf				
0209m_1.txt (6)	efforts to prevent or disrupt attacks and persistent vulnerability.				
	in describ <mark>in</mark> g America's prospects for civil liberties, Thieme warns that "The convergence of enabl <mark>in</mark> g technologies of intrusion, interception, and panoptic reach, comb <mark>in</mark> ed with a sense of urgency about do <mark>in</mark> g counter-terror and a clear				
	mandate from the White House to do everyth <mark>in</mark> g possible and seek forgiveness afterward rather than permission <mark>in</mark> adva has created a dire but often <mark>in</mark> visible set of threaten <mark>in</mark> g conditions."				
Options <<<	nas created a dire but often <mark>in</mark> visible set of threaten <mark>ing</mark> conditions.				
☐ scores ☐ highlight ☐ Filename ☐ highl word	"Concept of \$100 Laptop for World's Poor Is a Winner" USA Today (02/09/05) P. 3B; Maney, Kevi MIT Media Lab founder Nicholas Negroponte's unveiling of a \$100 laptop concept at last week's World Economic Forum spotlighted a workable idea for delivering advanced communications technology to people in hostile, poverty-plagued				
Palette <<<	regions. The laptop concept boasts more affordable display technology, a free operat <mark>in</mark> g system based on L <mark>in</mark> ux, 1 GB of				
	solid-state memory, and cheaper batteries that can be recharged with a hand crank. Several objectives are driving the push to bring technology and the Internet to the world's poor.				
high low	They range from the billions of new customers tech companies stand to ga <mark>in</mark> to an erosion of anti-American sentiment a				
Less_Water ▼	develop <mark>in</mark> g nations' economies, liv <mark>in</mark> g standards, educational systems, and cultures prosper from <mark>In</mark> ternet penetration. Negroponte's Hundred Dollar Laptop Corp. project has drawn back <mark>in</mark> g from major companies (AMD, Google), researchi (Hewlett-Packard's Alan Kay), and celebrities (Qu <mark>in</mark> cy Jones, Bono).				
	Meanwhile, the Our Economy nonprofit is work <mark>in</mark> g to br <mark>in</mark> g the <mark>in</mark> ternet <mark>in</mark> to low-income households, and AMD is roll <mark>in</mark> g o				
Search <<<	its \$185 Personal <mark>in</mark> ternet Communicator, a small device that lacks a keyboard and screen. Kev <mark>in</mark> Maney suggests that President Bush's recently released federal budget should commit fund <mark>in</mark> g to a "Bits of Americ				
Search word [in]	program for spreading technology to lower-income world populations.				
	"Chip That Speaks Languages of Asia Levels Playing Field" Wall Street Journal (02/09/04) P. B8; Ramstad, Evan				
Close	A breakthrough in microchip technology could significantly lower the cost of PCs in Ch <mark>in</mark> a and help remove Asia's				

Evaluation

Data Sets

We used three corpora for Evaluation:

- ACM Corpus: around 10000 articles
- Banksearch Corpus: around 10000 articles, topics about Banking and Finance, Programming Languages, Science and Sport
- Lyrics Corpus: 8000 lyrics texts, over 20 genres

Judges

Use of 5 judges for subjective and objective Evaluation to compare human and automatic scoring

Subjective Evaluation:

Use of questionnaires:

- about informativeness, readability & cohesion
- Likert scale

Comparison of automatic summaries with human written summaries:

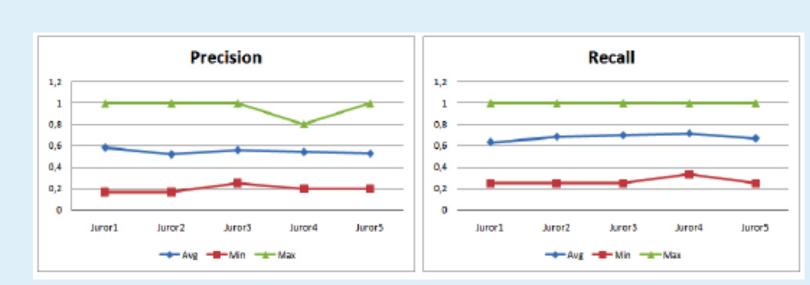
- grading(1-5) by other judges; results (average grades, placing):

Judges	System	Juror 1	Juror 2	Juror 3	Juror 4	Juror 5
Juror 1	3,0 (2)		2,1 (1)	3,1 (3)	3,2 (4)	3,8 (5)
Juror 2	4,0 (5)	2,6(2)		2,0 (1)	2,7(3)	3,4(4)
Juror 3	3,1 (3)	3,2(4)	3,3 (5)		2,8 (1)	2,9 (2)
Juror 4	3,9(5)	3,0 (3)	2,8 (2)	2,1(1)		3,2(4)
Juror 5	4,0 (5)	3,2(3)	1,8 (1)	2,5(2)	3,5(4)	
Overall	3,6 (6)	2,975(3)	2,45(2)	2,425(1)	3,05(4)	3,325(5)

Objective Evaluation:

Comparison of manually composed extracts:

- Precision & Recall Method



Comparison of extracts among each other:

- Inter-Indexer consistency: system has 59,8% (Rolling) and 60,3%(Cosine) average consistency

	System	Juror 1	Juror 2	Juror 3	Juror 4	Juror 5	Avg	Cosine Avg
System		60,3	57,5	61,4	60,8	59,2	59,8	60,3
Juror 1	60,3		71,2	68,3	62,1	69,2	66,2	66,3
Juror 2	57,5	71,0		72,3	68,8	66,8	67,3	67,4
Juror 3	61,4	68,3	72,3		69,8	66,8	67,7	67,8
Juror 4	60,8	62,1	68,8	69,8		84,2	69,1	66,9
Juror 5	59,2	69,2	66,8	66,8	84,2		69,2	67,0
						Overall	66,6	66,0

Significance Test

Wilcoxon-Test

Future Work

- multilingual summarization inclusion of images
- generation of abstracts