

Christa Womser-Hacker

Der PADOK-Retrievaltest

Zur Methode und Verwendung
statistischer Verfahren bei der Bewertung
von Information-Retrieval-Systemen

Technische Universität Darmstadt FACHBEREICH INFORMATIK	
B I B L I O T H E K	
Inventar-Nr.:	<u>MoS-00428</u>
Sachgebiete:	_____
Standort:	_____

1989

Georg Olms Verlag Hildesheim · Zürich · New York

Inhaltsverzeichnis

Vorwort

0. Einleitung	1
1. Der Kontext der Arbeit	3
1.1. Linguistische Informationswissenschaft	3
1.2. Das Projekt PADOK	3
1.2.1. Ausgangssituation	4
1.2.2. Komponenten der Evaluierung	5
1.2.3. Der PADOK-Retrievaltest	5
1.2.3.1. Organisatorischer Rahmen	5
1.2.3.2. Komponenten des Retrievaltests	6
1.2.3.2.1. Testdatenbanken und Dokumentauswahl	6
1.2.3.2.2. Aufgaben	6
1.2.3.2.3. Testpersonen	7
1.2.3.2.4. Die Relevanzbewertung	7
1.2.4. Ergebnisse von PADOK	9
1.2.4.1. Datenanalyse der Texterschließung	9
1.2.4.2. Statistische Auswertung des Retrievaltests auf der Grundlage von recall und precision	9
1.2.4.3. Analytische Auswertung der Retrievalprotokolle	10
1.2.4.4. Aufwandsmessung	10
1.3. Die Patentdokumentation	11
1.3.1. Patentdatenbanken	11
1.3.2. Recherchearten im Patentbereich	12
2. Information Retrieval	13
2.1. Terminologische Grundlagen und Definitionen	13
2.2. Information-Retrieval-Modelle	15
2.2.1. Wozu ein Modell?	15
2.2.2. Allgemeines Information-Retrieval-Modell	16
2.2.3. Boolesches Retrieval-Modell	16
2.2.4. Probabilistisches Retrieval-Modell	17
2.2.5. Vektorraum-Modell und Clustering-Ansatz	18
2.2.6. Fuzzy Set Theory	19
2.2.7. Kombinierte Modelle	19
2.2.8. Künstliche Intelligenz und Information Retrieval	19
2.3. Information Retrieval und Evaluierung	19
2.3.1. Die Problematik des Messens im Information Retrieval	20
2.3.2. Experiment und Untersuchung als Grundlagen der Evaluierung	21

2.3.3. Der Retrievaltest als Bewertungsinstrument	22
2.3.3.1. Ziele von Retrievaltests	22
2.3.3.2. Wesentliche Faktoren des Retrievaltests	22
2.3.3.3. Architektonischer Aufbau von Retrievaltests aus experimenteller Sicht	23
2.3.3.4. Exemplarische Darstellung von Retrievaltests	24
2.3.3.4.1. CRANFIELD 2	24
2.3.3.4.2. SMART	24
2.3.3.4.3. MEDLARS	25
2.3.3.4.4. Retrievaltests innerhalb des Projekts LIVE	25
2.3.3.4.5. Der AIR-Retrievaltest	25
2.4. Fazit	26
3. Effektivitätsmessung	27
3.1. Relevanz	27
3.1.1. Der Begriff der Relevanz	27
3.1.2. Die Relevanzproblematik	30
3.1.3. Systemrelevanz vs. Benutzerrelevanz	31
3.2. Elementarparameter zur Effektivitätsbewertung	32
3.3. Überblick über die wichtigsten Effektivitätsmaße	34
3.3.1. Die Standardmaße recall und precision	34
3.3.1.1. Der recall	35
3.3.1.2. Die precision	36
3.3.1.3. Einflußfaktoren bzgl. recall und precision	36
3.3.2. Weitere Maße	37
3.3.2.1. Der fallout	37
3.3.2.2. Komplementmaße zu recall, precision und fallout: miss ratio, noise ratio und rejection ratio	38
3.3.2.3. Resolution factor und elimination factor	38
3.3.2.4. Die Generality	39
3.3.3. Komplexe Maße	40
3.3.3.1. Koordination von Maßen	41
3.3.3.1.1. Halbordnung/Ordnung	41
3.3.3.1.2. Graphische Kombinationsmaße	41
3.3.3.1.2.1. Der precision-recall-Graph	41
3.3.3.1.2.2. Der fallout-recall-Graph	44
3.3.3.1.3. Punktwolken	44
3.3.3.2. Maße für Ranking-Verfahren und mehrstufige Relevanzskalen	45
3.3.3.3. „Single number measures“	46
3.3.3.3.1. Einfache Verknüpfungen der Elementarparameter	46
3.3.3.3.2. Die Maße von Swets und Brookes	46
3.3.3.3.3. Das Heine-Maß	47
3.3.3.3.4. Meetham's I-Maß	48
3.3.3.3.5. Das e-Maß Van Rijsbergens	48
3.4. Zusammenfassung	54
3.5. Prinzipien der Maßauswahl	55

3.5.1. Das Retrievalmodell	55
3.5.2. Benutzerstandpunkt	55
3.5.3. Meßskala	56
3.5.4. Verfahren zur Bewertung von Maßen	57
3.5.4.1. Meßtheoretischer Ansatz	57
3.5.4.2. Empirisch-pragmatischer Ansatz	57
3.6. Der Sonderfall Null-durch-Null oder Logarithmus von Null	58
3.7. Fazit: Effektivitätsmaße	58
4. Beschreibung von Verteilungen auf der Grundlage von Mittelwerten, Streuungswerten und Korrelationsmaßen	59
4.1. Empirische Verteilungen	59
4.2. Univariable Verteilungen	59
4.2.1. Verfahren der Mittelwertbildung	60
4.2.2. Auswahlkriterien und Vergleich der Mittelwerte	60
4.2.2.1. Verteilungsformen	61
4.2.2.2. Skalierung	62
4.2.2.3. Vergleich der Mittelwerte	64
4.2.3. Streuungsmaße	65
4.2.4. Mittelwerte bei der Bewertung von Retrievalergebnissen	67
4.2.4.1. Makro- und Mikromittelung	67
4.2.4.2. Inhaltliche Differenzierung zwischen Makro- und Mikromethode	68
4.2.4.3. Axiome der Mittelwertbildung	68
4.2.4.4. Fazit: Mittelwert und Streuung	69
4.3. Bivariable Verteilungen	69
4.3.1. Die graphische Darstellung bivariabler Verteilungen	70
4.3.2. Korrelationsmaße	71
4.3.2.1. Intervallskalierte Variablen	71
4.3.2.2. Ordinalskalierte Variablen	72
4.4. Fazit	73
5. Statistische Signifikanz	74
5.1. Grundlagen der Prüfstatistik	74
5.1.1. Stichprobentheorie	74
5.1.1.1. Grundgesamtheit, Stichprobe und Zufallsvariable	74
5.1.1.2. Funktion von Stichproben	75
5.1.1.3. Problem der Repräsentativität und Stichprobenfehler	75
5.1.1.4. Größe von Stichproben	77
5.1.1.5. Stichprobenbildung und Randomisierung	78
5.1.2. Grundlagen der Wahrscheinlichkeitsrechnung	80
5.1.2.1. Der Begriff der Wahrscheinlichkeit	80
5.1.2.2. Axiome der Wahrscheinlichkeitsrechnung	80
5.1.2.3. Kombinatorik	81
5.1.2.4. Punkt- und Überschreitungswahrscheinlichkeiten	82

5.1.2.5. Wahrscheinlichkeitsverteilungen	82
5.1.2.5.1. Die Binomialverteilung	82
5.1.2.5.2. Die Normalverteilung	84
5.1.2.5.2.1. Das statistische Modell der Normalverteilung	84
5.1.2.5.2.2. Eigenschaften der Normalverteilung	85
5.2. Die Überprüfung der statistischen Signifikanz	87
5.2.1. Allgemeiner Ablauf von Signifikanztests	87
5.2.2. Prinzipien der Testauswahl	89
5.2.2.1. Parametrische vs. nichtparametrische Verfahren	90
5.2.2.1.1. Gegenüberstellung	90
5.2.2.1.2. Stichprobenumfang und Effizienz	91
5.2.2.2. Tests für zwei oder k Stichproben	92
5.2.2.2.1. Tests für zwei Stichproben	92
5.2.2.2.1.1. Tests für zwei abhängige Stichproben	92
5.2.2.2.1.2. Tests für zwei unabhängige Stichproben	94
5.2.2.2.2. Tests für k Stichproben	95
5.2.2.2.2.1. k abhängige Stichproben	95
5.2.2.2.2.2. k unabhängige Stichproben	95
5.2.2.2.3. Vergleichende Gegenüberstellung der Tests	96
5.2.2.3. Stichprobenumfang	97
5.2.2.4. Stärkeeffizienzvergleich	98
5.2.2.5. Fazit: Testauswahl	98
6. Bewertung der Retrievalergebnisse des PADOK-Retrievaltests	100
6.1. Beschreibung der Materialgrundlage	100
6.2. Vergleich der Antwortmengen	101
6.2.1. Die Gesamtantwortmenge	101
6.2.2. Differenzierung der Dokumente nach Qualität	103
6.2.2.1. Die relevanten Dokumente	104
6.2.2.2. Der Ballast	104
6.2.3. Die inhaltliche Überschneidung	106
6.2.3.1. Die inhaltliche Überschneidung bei den relevanten Dokumenten	106
6.2.3.2. Die inhaltliche Überschneidung bei den Ballastdokumenten	108
6.2.3.3. Fazit: Inhaltliche Überschneidung	109
6.3. Statistische Auswertung auf der Grundlage von Bewertungsmaßen	110
6.3.1. Sequenzanalyse der Stichproben der recall- und precision-Werte	110
6.3.2. Makrorecall und Makroprecision	112
6.3.3. Mikrorecall und Mikroprecision	113
6.3.4. Recall- und precision-Mittelwerte auf der Grundlage des Medians	113
6.3.5. Zweistufiges Ballastkonzept	114
6.3.6. Die Streuung der Meßwerte	115
6.3.6.1. Die Streuung der recall-Werte	115
6.3.6.2. Die Streuung der precision-Werte	118
6.3.7. Überprüfung der Korrelation zwischen recall und precision	119
6.3.8. Recall-precision-Paare	121

6.3.8.1. Punktwolken	121
6.3.8.2. Vergleich der Retrievalergebnisse auf der Basis von recall-precision-Paaren	122
6.3.8.3. Fazit: recall-precision-Paare	124
6.3.9. Anwendung komplexer Maße	124
6.3.9.1. Das e -Maß nach Van Rijsbergen	125
6.3.9.2. Das I -Maß von Meetham	125
6.3.9.3. e -Maß vs. I -Maß	126
6.3.9.4. e -Maß vs. heuristisches Maß	127
6.3.10. Fazit: Bewertung auf der Grundlage von Maßen	128
6.4. Die Kontrolle der Randbedingungen mittels Blockbildung	130
6.4.1. Blockbildung nach Prüfergruppen	130
6.4.1.1. Antwortmengen	131
6.4.1.2. Makroberechnung von recall, precision und e -Maß	133
6.4.1.3. Fazit: Blockbildung nach Prüfergruppen	134
6.4.2. Blockbildung nach Testpersonen	134
6.4.2.1. Antwortmengen	135
6.4.2.2. Makroberechnung von recall, precision und e -Maß	138
6.4.2.3. Fazit: Blockbildung nach Testpersonen	140
6.4.3. Blockbildung nach Testpersonen und Testphasen	141
6.4.4. Blockbildungen nach Größe der Antwortmenge und Anzahl der Teilsuchprozesse	144
6.4.5. Blockbildung nach der Größe der Generality	146
6.4.6. Blockbildung zur Kontrolle von Eigenschaften der Aufgaben	147
6.4.6.1. Differenzierung der Aufgaben nach IPC-Unterklassen	147
6.4.6.2. Differenzierung der Aufgaben nach dem Vorhandensein einer Zeichnung	148
6.4.7. Fazit: Blockbildungen	150
6.5. Die Problematik der Nullantworten	151
6.5.1. Definition von Nullantwort	151
6.5.2. Nullantworten und Bewertungsmaße	151
6.5.3. Verteilung der Nullantworten	152
6.5.3.1. Verteilung der Nullantworten auf die Prüfergruppen	153
6.5.3.2. Verteilung der Nullantworten auf Testpersonen	153
6.5.3.3. Verteilung der Nullantworten auf Testphasen	154
6.5.4. Differenzierung der Nullantworten nach Ballast	155
6.5.5. Nullantworten und Teilsuchprozesse	156
6.5.6. Fazit: Nullantworten	156
6.6. Anwendung von Signifikanztests	157
6.6.1. Problemstellung, Grundlage und Nullhypothese	157
6.6.2. Vorzeichentest	157
6.6.2.1. Methode des Vorzeichentests	158
6.6.2.2. Ergebnisse und Interpretation	158
6.6.2.3. Fazit: Vorzeichentest	159
6.6.3. Wilcoxon-Vorzeichenrang-Test	160
6.6.3.1. Testablauf und Methode	160
6.6.3.2. Ergebnis des Wilcoxon-Vorzeichenrang-Tests	161

6.6.3.3. Fazit: Wilcoxon-Vorzeichenrang-Test	162
6.6.4. Zwei-Weg-Rangvarianzanalyse von Friedman	162
6.6.4.1. Testablauf und Methode	162
6.6.4.2. Ergebnis der Zwei-Weg-Rangvarianzanalyse bzgl. recall	163
6.6.4.2.1. Test über die Gesamtmenge der Aufgaben	163
6.6.4.2.2. Blockbildungen nach Prüfergruppen und Testphasen	163
6.6.4.3. Ergebnis der Zwei-Weg-Rangvarianzanalyse bzgl. precision	165
6.6.4.3.1. Test über die Gesamtmenge der Aufgaben	165
6.6.4.3.2. Blockbildungen nach Prüfergruppen und Testphasen	165
6.6.4.4. Fazit: Zwei-Weg-Rangvarianzanalyse nach Friedman	167
6.6.5. Fazit: Signifikanzüberprüfung	167
7. Zusammenfassung der Ergebnisse	168
7.1. Das Anwendungsgebiet der Patentedokumentation	168
7.2. Statistische Gesamtbewertung	168
7.3. Die Rolle der Statistik	171
ANHANG	
A: Beispiel für eine Aufgabe des PADOK-Retrievaltests	172
B: Streuung der precision-Werte des PADOK-Retrievaltests	173
Literaturverzeichnis	174