Zürcher Hochschule
für Angewandte Wissenschaften

Zürcher Hochschule für angewandte Wissenschaften ZHAW

SPEZIELLE HERAUSFORDERUNGEN BEI DER EVALUATION DER INHALTSERSCHLIESSUNG

Martin Braschler

Prof. ZFH, Dr. sc.

Stv. Leiter Institut für angew. Informationstechnologie

Leiter Forschungsgruppe «Information Engineering»

Zur Person

Ausbildung/Berufliches:

- Dr. sc., Dipl. Informatik-Ing. ETH
- Professor ZFH, Dozent
- Stv. Leiter Institut für angewandte Informations-technologie InIT, Zürcher Hochschule für ange-wandte Wissenschaften ZHAW, Winterthur

Wissenschaftliche Themen:

- Experte für Information Retrieval (IR)
- Schwerpunkte Evaluation von IR, mehrsprachiges IR
- Mitgründer der CLEF-Evaluationskampagne für IR
- Deputy Steering Committee Chair CLEF Association

Forschungsfragen

- Was können wir lernen aus dem reichen Erfahrungsschatz der (experimentellen) Information Retrieval-Evaluation?

- Welche Sacherschliessungsmethode ist
  (1) gut automatisierbar, ***und***
  (2) bietet den Nutzern bessere Suchresultate?

Sacherschliessung als Mittel zum Zweck (Suche)

Abgrenzung:

- Im Kontext dieser Präsentation sind wir an der Frage interessiert, inwiefern die Sacherschliessung eine bessere Retrievaleffektivität («bessere Suchresultate») ermöglicht im Vergleich zu einem Volltextretrieval ohne (manuelle) Erschliessung

- Die Messung erfolgt also indirekt – nicht die Sacherschliessung selbst wird beurteilt, sondern ihr Beitrag zu den Suchresultaten
  (nicht die «*beste Sacherschliessung*», sondern die «*beste Suche dank der Sacherschliessung*»)

Evaluation von Retrievaleffektivität

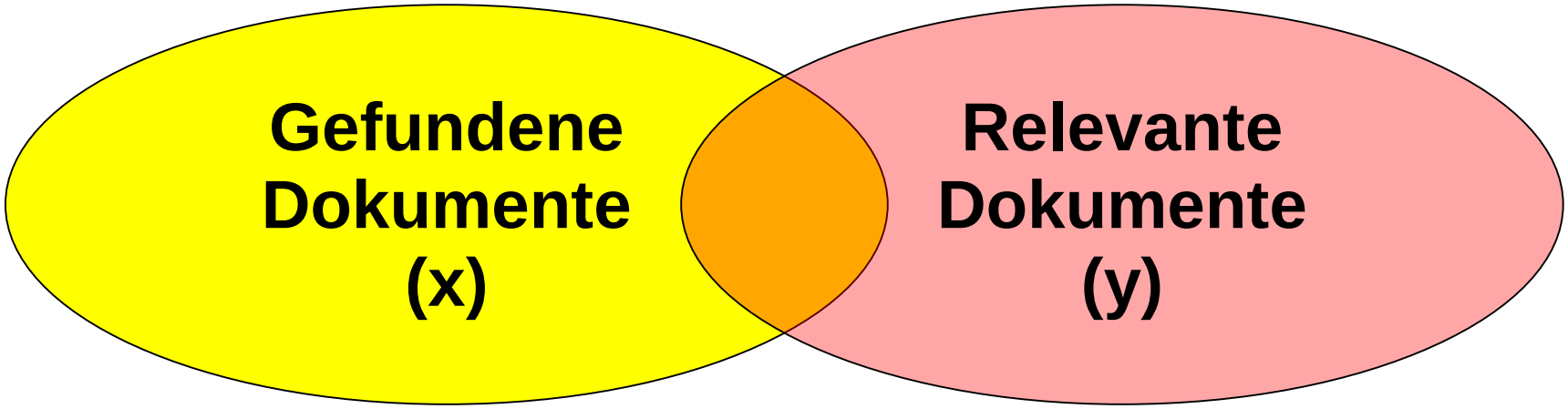
Evaluation von Retrievaleffektivität («Suchqualität») erfolgt populärerweise durch die Masse Ausbeute und Präzision (Voorhees, 2001).

Das Ziel: «Auffinden von möglichst viel relevanter Information bei gleichzeitiger Minimierung der ebenfalls gelieferten irrelevanten Information» («Retrievalproblem»)

Ausbeute = «möglichst viel relevante Information»
Präzision = «Minimierung der ebenfalls gelieferten irrelevanten Information»

Präzision und Ausbeute

Präzision = (x ∩ y)/x
Ausbeute = (x ∩ y)/y



Aber: die beiden Masse stehen im Widerspruch!
Hohe Präzision → niedrige Ausbeute
Hohe Ausbeute → niedrige Präzision

Arbeiten aus der Frühzeit des IR

IR hat sich schon früh mit der Frage befasst, ob Volltextsuche eine manuelle Indexierung/Verschlagwortung ersetzen kann.

Die klassische Sichtweise aus diesen Experimenten war:

Manuelle Erschliessung, kontrolliertes Vokabular = gut für Präzision
Automatische Erschliessung, Volltextretrieval = gut für Ausbeute*

* Es gab Indikationen, dass vor allem Fachpersonen mit manueller Erschliessung auch für Ausbeute bessere Ergebnisse erzielen («OHSUMED Experiments», zitiert nach Savoy 2004)

Eine moderne Sicht auf das Problem

Aus heutiger Sicht sollte der Fokus aber ein anderer sein:

1. Wie können sich die beiden Paradigmen ergänzen, und

2. Welche Sacherschliessungsmethoden sind so automatisierbar, dass dies im grossen Stil umgesetzt werden kann?

Durchführung einer IR-Evaluation

Die Evaluation erfolgt nach dem Cranfield-Paradigma

Grundlage ist eine Testkollektion:

- eine Liste von Informationsbedürfnissen fiktiver Nutzer,
- eine statische Dokumentenkollektion,
- und Relevanzbewertungen

Der Bau einer solchen Testkollektion ist sehr aufwändig und teuer (u.a. |D| x |Q| Relevanzbewertungen), und daher für spezifische Dokumentenkollektion kaum praktikabel

Mögliche Vorgehensweisen für die Evaluation unter Einbezug einer Sacherschliessung

1. Verwendung einer Testkollektion mit sowohl (Voll-)texten als auch Deskriptoren aus einem kontrollierten Vokabular
   - Beispiel: Amaryllis (aber: nur Abstract)

2. Anreicherung einer bestehenden Testkollektion mit Volltexten mit automatisch generierten Deskriptoren (Domäne muss passen)

- Szenario 1 kann um Szenario 2 erweitert werden (direkter Vergleich manuelle vs. automatische generierte Deskriptoren)

Messung vs. Retrievalparadigmen

Messung der Retrievaleffektivität unter Einbezug der Deskriptoren

Schwierigkeiten:

- Volltextretrieval und die Suche mittels Deskriptoren eignen sich teilweise für unterschiedliche Retrievalparadigmen
  - Gewichtetes Retrieval mit rangierten Resultaten vs.
  - Boole’sches Retrieval mit Resultatmengen

Vorgehen beim Volltextretrieval

Die Gewichtung der Suchbegriffe im Volltextretrieval stützt sich grundsätzlich darauf, wie «charakteristisch» Begriffe für ein Dokument sind

Beispiel: tf*idf: lokale Häufigkeit mal globale Seltenheit

Weiter kann beispielsweise ermittelt werden, wie gut ein Begriff zwischen relevanten und irrelevanten Dokumenten diskriminiert

Eine Grundannahme hier ist, dass Auftretenshäufigkeiten hierbei eine Aussage zur Relevanz zulassen

Vergleich Tokens aus Volltext mit Deskriptoren

Es ist unklar, ob wir manuell (oder auch automatisch) zugeordnete Deskriptoren auf die gleiche Weise behandeln sollen, resp. ob diese in Sachen Häufigkeiten den gleichen Gesetzmässigkeiten folgen (so gilt z.B. für Deskriptoren im Allgemeinen immer tf=1!)

Auf der anderen Seite ist Boole’sches Retrieval, mit einer scharfen Unterscheidung «vorhanden/nicht vorhanden» bei offenem Vokabular nicht geeignet

Bestehende Messungen

Die Messung der Retrievaleffektivität unter Einbezug der Deskriptoren wurde teilweise bereits untersucht.

Die Resultate sind gemischt. Es gibt wenig Hinweise, dass manuelle Erschliessung bessere Retrievaleffektivität als Volltextsuche bringt (Resultate aus Studien der Frühzeit des IR, zitiert nach Savoy 2004)

Aber: es gibt starke Hinweise, dass eine «Combination of Evidence» interessant ist! Siehe auch die Resultate von Savoy (2004) auf der Amaryllis-Kollektion.

Amaryllis-Kollektion

<DOC>
<DOCNO> AM-000004 </DOCNO>
<TI> Les marchés de l'environ créent plus d'emplois que de métiers </TI>
<AB> A mesure que l'observation du marché de l'emploi environnement se développe, les tendances enregistrées depuis quelques années se confirment. Des emplois en augmentation régulière mais des professions et des métiers encore peu nombreux et peu reconnus, une relation formation-emploi difficile à trouver, des métiers écartelés entre faibles et hautes qualifications: les décalages du marché de l'emploi environnement sont encore importants. Il n'en reste pas moins que les préoccupations d'environnement semblent avoir trouvé leur place sur le marché de l'emploi: plutôt que "vague verte", l'environnement s'inscrit dans la durée </AB>
<MC> Protection environnement, Emploi, Marché travail </MC>
<KW> Environmental protection, Employment, Labour market </KW></DOC>

Für gewichtetes Retrieval:

Für Average Precision
- Kombination > Manuelle Deskriptoren > Abstract + Titel

Für Präzision @ 5, 10, 20
- Kombination > Manuelle Deskriptoren = Abstract + Titel

Einschränkungen dieser Evaluation

Herausforderungen:

- Viele verschiedene Gewichtungsformeln – wurden die richtigen verwendet?
- Ist dieses Resultat auch gültig für Volltext statt Abstract und Titel?
- Wurden die richtigen Evaluationsmasse verwendet?
  - Average Precision vermischt die Aspekte «Ausbeute» und «Präzision»
  - Präzision @ x ist ein instabiles Mass

Herausforderungen für die Evaluation

- Zusammenspiel aus Qualität der Sacherschliessung mit dem Retrievalresultat bleibt unbekannt
- Wie wird kombiniert? Sind Deskriptoren einfach «weitere Tokens»? Gleiche Gewichtung?
- Wie werden evtl. Teilscores aus beiden Teilen kombiniert?
- Was wären die richtigen Masse?

Fragen

- Wir können wir die vorhandenen Erfahrungen und die Infrastruktur der (experimentellen) IR-Evaluation nutzen?
  - Verschlagworten einer IR-Testkollektion
  - Relevanzbewertungen für verschlagwortete Kollektionen
- Wie muss eine Testkollektion beschaffen sein? Welche Domänen soll sie abdecken?
- Wie muss die Sacherschliessung angepasst werden für eine bessere Retrievaleffektivität? (Anzahl Deskriptoren, Granularität, …)
- Welches Erschliessungsystem eignet sich für welchen Einsatz? (präzisions-orientiert, ausbeute-orientiert, etc.)

Quellen

Voorhees, E. M. (2001, September). The philosophy of information retrieval evaluation. In *Workshop of the cross-language evaluation forum for european languages* (pp. 355-370). Springer, Berlin, Heidelberg.

Savoy, J. (2005). Bibliographic database access using free-text and controlled vocabulary: an evaluation. *Information processing & management*, *41*(4), 873-890.