Applying Rough Set Theory to Information Retrieval

CS836 – Class Presentation
Instructor: Dr. Wojtek Ziarko

Bing Zhou
10/29/2008

1

Outline

1. Motivation
2. Background: Information retrieval (IR) theories and technologies
3. Issues with existing IR models
4. A rough set (RS) approach to IR
5. Other applications
6. Project Goal & Status
7. Conclusion
8. References

2

Motivation

1. To find connections between RS and IR by classifying existing approaches to this topic
2. To find possible improvements of the existing solutions
3. To find more applications for employing RS in IR

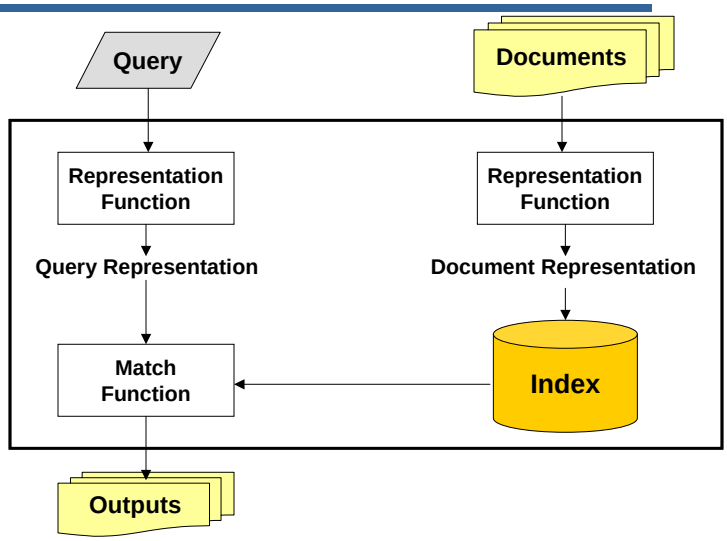
3

Background:What is IR?

1. IR is a branch of information processing which aims at searching documents from the database suitably for the user's interest
2. There are numerous bibliographical data of documents stored in huge databases and the number of documents is growing day after day
3. An ideal IR system which provides the list of appropriate documents to the user is desirable but it is very difficult to construct such a system
4. Are we satisfied with Google?

4

The IR Work Flow



5

Binary Representation of Documents

Document 1

The quick brown fox jumped over the lazy dog's back.

Document 2

Now is the time for all good men to come to the aid of their party.

Term	Document 1	Document 2
aid	0	1
all	0	1
back	1	0
brown	1	0
come	0	1
dog	1	0
fox	1	0
good	0	1
jump	1	0
lazy	1	0
men	0	1
now	0	1
over	1	0
party	0	1
quick	1	0
their	0	1
time	0	1

Stopword List

for
is
of
the
to

6

How do We Weight Document Terms?

- Here's the intuition:
  - Terms that appear often in a document should get high weights

    The more often a document contains the term "dog", the more likely that the document is "about" dogs.

  - Terms that appear in many documents should get low weights

    Words like "the", "a", "of" appear in (nearly) all documents.
- How do we capture this mathematically?
  - Term frequency
  - Inverse document frequency

7

Weighted Representation of Documents

$w_{i,j}$ **weight assigned to term *i* in document *j***

$\mathrm{tf}_{i,j}$ **number of occurrence of term *i* in document *j***

N **number of documents in entire collection**

n_i **number of documents with term *i***

$$w_{i,j} = \mathrm{tf}_{i,j} \cdot \log \frac{N}{n_i}$$

Relative importance inside a document

Relative importance in a collection

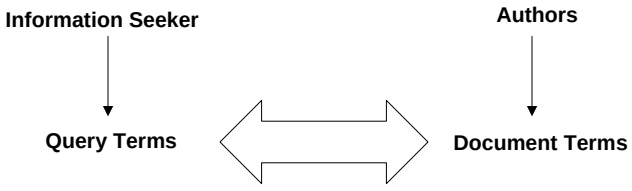
8

TF.IDF Example

	TF				IDF	$W_{i,j}$			
	1	2	3	4		1	2	3	4
complicated			5	2	0.301			1.51	0.60
contaminated	4	1	3		0.125	0.50	0.13	0.38	
fallout	5		4	3	0.125	0.63		0.50	0.38
information	6	3	3	2	0.000				
interesting		1			0.602		0.60		
nuclear	3		7		0.301	0.90		2.11	
retrieval		6	1	4	0.125		0.75	0.13	0.50
siberia	2				0.602	1.20			

9

The Central Problem in IR



Do they match?
Do these represent the same concepts (relevance level)?

10

Different Models of IR

1. Boolean model
2. Fuzzy set model
3. Vector space model
4. Probabilistic model

11

Boolean Model

- Based on a simple theory-Boolean logic, easy to understand
- Both the query and documents are indexed by a set of terms
- Documents are retrieved only if they exactly match the Boolean conditions specified in the query

12

Boolean Retrieval

- Weights assigned to terms are either "0" or "1"
  - "0" represents "absence": term **isn't** in the document
  - "1" represents "presence": term **is** in the document
- Build queries by combining terms with Boolean operators
  - AND, OR, NOT
- The system returns all documents that satisfy the query

13

Logic Tables

A \ B	0	1
0	0	1
1	1	1

A OR B

B	0	1
	1	0

NOT B

A \ B	0	1
0	0	0
1	0	1

A AND B

A \ B	0	1
0	0	0
1	1	0

A NOT B
(= A AND NOT B)

14

Boolean View of a Collection

Term	Doc 1	Doc 2	Doc 3	Doc 4	Doc 5	Doc 6	Doc 7	Doc 8
aid	0	0	0	1	0	0	0	1
all	0	1	0	1	0	1	0	0
back	1	0	1	0	0	0	1	0
brown	1	0	1	0	1	0	1	0
come	0	1	0	1	0	1	0	1
dog	0	0	1	0	1	0	0	0
fox	0	0	1	0	1	0	1	0
good	0	1	0	1	0	1	0	1
jump	0	0	1	0	0	0	0	0
lazy	1	0	1	0	1	0	1	0
men	0	1	0	1	0	0	0	1
now	0	1	0	0	0	1	0	1
over	1	0	1	0	1	0	1	1
party	0	0	0	0	0	1	0	1
quick	1	0	1	0	0	0	0	0
their	1	0	0	0	1	0	1	0
time	0	1	0	1	0	1	0	0

Each column represents the view of a particular document: What terms are contained in this document?

Each row represents the view of a particular term: What documents contain this term?

To execute a query, pick out rows corresponding to query terms and then apply logic table of corresponding Boolean operator

15

Sample Queries

Term	Doc 1	Doc 2	Doc 3	Doc 4	Doc 5	Doc 6	Doc 7	Doc 8	
dog	0	0	1	0	1	0	0	0	
fox	0	0	1	0	1	0	1	0	
dog $\wedge$ fox	0	0	1	0	1	0	0	0	dog AND fox $\rightarrow$ Doc 3, Doc 5
dog $\vee$ fox	0	0	1	0	1	0	1	0	dog OR fox $\rightarrow$ Doc 3, Doc 5, Doc 7
dog $\neg$ fox	0	0	0	0	0	0	0	0	dog NOT fox $\rightarrow$ empty
fox $\neg$ dog	0	0	0	0	0	0	1	0	fox NOT dog $\rightarrow$ Doc 7

16

Strengths and Weaknesses of the Boolean Model

- Strengths
  - Precise, if you know the right strategies
  - Precise, if you have an idea of what you're looking for
  - Efficient for the computer
- Weaknesses
  - Users must learn Boolean logic
  - Boolean logic insufficient to capture the richness of language
  - No control over size of result set: either too many documents or none
  - When do you stop reading? All documents in the result set are considered "equally good", no ranking provided
  - What about partial matches? Documents that "don't quite match" the query may be useful also

17

The Needs for Ranked Retrieval

- Too many resources in the database, the user often get more results than the amount they are willing to examine
- Some documents are more relevant than others, relevance has degree

18

Similarity-Based Models

- Let's replace relevance with "similarity"
  - Rank documents by their similarity with the query
- Treat the query as if it were a document
- Find its similarity to each document
- Rank the documents by similarity

19

Fuzzy Set Model

- Fuzzy Set Theory
  - Extension of Boolean set theory
  - Instead of a binary membership definition, fuzzy set Membership is continuously defined between 0 and 1
  - Example:
    - { Male students in our class}
    - {tall students in our class}
    - One is Boolean set and one is fuzzy set

20

Fuzzy Set Model

- The set of retrieved documents should be considered as a fuzzy set
- Documents are not just relevant or non-relevant
- Documents can be somehow relevant
- Documents are ranked by their relevance degrees
- Good Mathematical Models but not widely implemented and tested

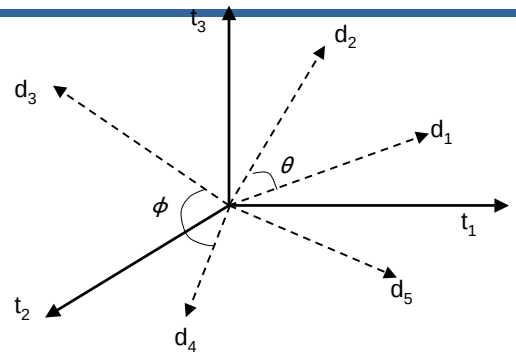
21

Vector Space Model

- Based on geometry, both query and document are defined as vectors
- Matching is done through a similarity measure (cosine measure)
- Documents are ranked based on their similarity to the query (ranked retrieval)
- Best/partial match

22

Vector Space Model



Matching: evaluate distance, documents that are "close together" in vector space considered similar

Therefore, retrieve documents based on how close the document is to the query (i.e., similarity ~ "closeness")

23

Similarity Measure

- use "angle" between the vectors:

$$\cos(\theta) = \frac{\vec{d}_j \cdot \vec{d}_k}{\left|\vec{d}_j\right|\left|\vec{d}_k\right|}$$

$$sim(d_j, d_k) = \frac{\vec{d}_j \cdot \vec{d}_k}{\left|\vec{d}_j\right|\left|\vec{d}_k\right|} = \frac{\sum_{i=1}^{n} w_{i,j} w_{i,k}}{\sqrt{\sum_{i=1}^{n} w_{i,j}^2}\sqrt{\sum_{i=1}^{n} w_{i,k}^2}}$$

24

Components of the Similarity Measure

- The "inner product" (numerator) is the key to the similarity function

$$\vec{d}_j \cdot \vec{d}_k = \sum_{i=1}^{n} w_{i,j} w_{i,k}$$

Example: $[1\ 2\ 3\ 0\ 2] \cdot [2\ 0\ 1\ 0\ 2]$

$$= 1\times 2 + 2\times 0 + 3\times 1 + 0\times 0 + 2\times 2 = 9$$

- The denominator handles document length normalization

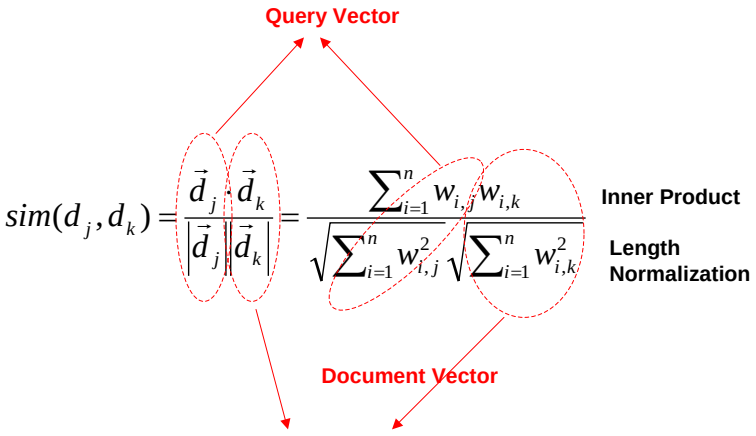
$$\left|\vec{d}_j\right| = \sqrt{\sum_{i=1}^{n} w_{i,k}^2}$$

Example: $\left|[1\ 2\ 3\ 0\ 2]\right|$

$$= \sqrt{1+4+9+0+4} = \sqrt{18} \approx 4.24$$

25

Reexamining Similarity



26

Summary:Key Ideas in VSM

- Goal: find documents most similar to the query
- Geometric interpretation: measure similarity in terms of angles between vectors in high dimensional space
- Documents weights can be computed by TF * IDF
- Length normalization is critical
- Similarity is calculated via the inner product

27

Probabilistic Model

- Most theoretically sound
- Attempts to estimate the probability that the User will find a given document relevant for their query
- Captures the IR problem within a probabilistic framework
- Document ranking by probabilities

28

Probabilistic Model

- Given a query *q* and a document *dj*
  - Probabilistic model attempts to estimate the probability that *dj* is relevant
  - Similarity between q and dj = odds of dj being relevant

$$Sim\ (q, d_j) = \frac{P\ (d_j\ relevant\ to\ q)}{P\ (d_j\ not\ relevant\ to\ q)}$$

29

The Trends of IR

1. From exactly matched unranked outputs to partially matched ranked list
2. Attempts to improve IR quality by using intelligent matching strategies, e.g., use concept analysis requiring semantic calculations, the RS approach is trying to address this problem

30

Issues with the Existing IR Models

1. The user is expected to specify terms that refer to the subject of interest, which is a big burden
2. The obtained recall level is low
3. The needs for more user oriented and flexible search strategies
4. … …

31

A Rough Set Approach to IR

1. "Approximation for Information Retrieval" by Padmini Srinivasan, one of the earliest approach of applying RS to IR
2. Main objective: using rough approximation to improve IR system performance

32

Main Problem-1

1. Variations in search vocabulary
   - Example: "education" related terms: university, student, faculty, school, study… …
2. Different user type choose to use different terms to do search
   - Beginners
   - Experts
3. Existing solutions
   - Design advanced front ends that can intelligently select terms
   - Thesaurus: http://www.snap.com/

33

Existing Solutions: An Artificial Intelligence Approach

1. Design an expert system that performs the job of information specialist
2. Goal: to design systems which share the responsibility of an effective search with the user

34

Main Problem-2

- To improve the recall level
  - Precision is defined as the proportion of retrieved documents that are actually relevant
  - Recall is defined as the proportion of relevant documents that are actually retrieved

35

Rough Approximation

1. Lower approximation
   - A set of concepts which has complete representation in the document/query
   - $\underline{R}X = \bigcup\{Y \in U / R : Y \subseteq X\}$
2. Upper approximation
   - A set of concepts which has at least a partial representation in the document/query
   - $\overline{R}X = \bigcup\{Y \in U / R : Y \cap X \neq \varnothing\}$

36

Rough Equality

- Two sets are approximately equal
  - Sets *X* and *Y* are bottom *R*-equal $(X \approx Y)$

  If $\underline{R}X = \underline{R}Y$
  - Sets *X* and *Y* are top *R*-equal $(X \simeq Y)$

  If $\overline{R}X = \overline{R}Y$
  - Sets *X* and *Y* are *R*-equal $(X \approx Y)$

  If $(X \approx Y)$ and $(X \simeq Y)$

37

Rough Inclusion

- One set is approximately included in another set
  - Sets *X* is bottom *R*-included in *Y* $(X \underset{\sim}{\subset} Y)$

  Iff $\underline{R}X \subseteq \underline{R}Y$
  - Sets *X* is top *R*-included in *Y* $(X \tilde{\subset} Y)$

  Iff $\overline{R}X \subseteq \overline{R}Y$
  - Sets *X* is *R*-included in *Y* $(X \tilde{\underset{\sim}{\subset}} Y)$

  Iff $(X \underset{\sim}{\subset} Y)$ and $(X \tilde{\subset} Y)$

38

The RS Approach-Main Methodologies

1. Make approximate matches between query and documents
2. Organizing terms into an approximation space of equivalence classes
3. Each class contains terms that are semantically equivalent
   - Example: corporation, organization, industry
4. The term classes are identified by subject experts

39

An Example - The Approximation Space

1. 10 terms that are partitioned into 4 classes
   - C1: t1,t2,t5
   - C2: t3,t7
   - C3: t4,t8,t9,t10
   - C4: t6
2. Document D with term {t1,t2,t6}
3. Lower approximation: C4
4. Upper approximation: C1 and C4

40

Approximation Matching Strategies – 4 steps

1. The document has all the terms specified in the query
   - Boolean model
2. The document approximately matches the query
   - Rough equality: bottom-equal: more relevant; top-equal: less relevant
3. Query concepts contained in the document or vice versa
   - Rough inclusion: bottom-included: more relevant; top-included: less relevant
4. Based on the overlap between document and query
   - bottom-overlap: more relevant; top-overlap: less relevant
5. Documents are ranked based on the above retrieval strategy

41

Advantage Comparing to Other IR Model

- Even if authors (of documents) and searchers (of query) use different terms, retrieval can still be effective because the underlying approximation space captures their semantic equivalence

42

Experimental Results

- Compare with VSM model
  - RS model provides better ranking
  - RS model provides significant improvement in recall
  - The vocabulary problem may be handled within the constructs of the model itself, without having an advanced front-end

43

Other Applications

1. "A Machine Learning Approach to Information Retrieval" by Wong and Ziarko
   - An interactive learning algorithm which leads to fast retrieval of the majority of relevant documents
2. "Optimal Determination of User-oriented Clusters" by Vijay V. Raghavan
   - Using a measure derived from rough sets to identify cluster boundaries

44

Project Goal

- Investigate the applications of RS theory in the field of IR
  - To collect and classify classical papers with regard to this topic
  - To look into the possible improvements of the solutions provided in these papers
  - To find out new applications for employing RS in IR
  - To address the fundamental connections between these two theories

45

Project Status

- I stared this work at the beginning of October, and I have finished the paper collection part, I am continuing the rest of the work until I have enough findings to conclude in the project report. All this will of course be much more detailed than it is in the proposal and this presentation, with things added to make the report complete and concise

46

Conclusion

- The existing IR models have some practical issues
  - The user is expected to specify terms that refer to the subject of interest
  - The obtained recall level is low
  - The needs for more user oriented and flexible search strategies
- Applying RS to IR, there is a possibility to:
  - Improve IR system performance
  - Relieve the search burden from the user

47

References

- Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, Introduction to Information Retrieval, Cambridge University Press. 2008
- Pawlak, Zdzisław, Rough Sets: Theoretical Aspects of Reasoning About Data. Dordrecht: Kluwer Academic Publishing. 1991
- Lecture Notes
- Padmini Srinivasan, Approximation for Information Retrieval, Information Processing and Management, 1989
- S. K. Michael Wong, Wojciech Ziarko , A Machine Learning Approach to Information Retrieval, Research and Development in Information Retrieval. 1986
- Vijay V. Raghavan, Jitender S. Deogun, Optimal Determination of User-Oriented Clusters. SIGIR. 1987

48

Questions?

49