Information Retrieval

Lecture 1: Introduction to concepts and problems

Computer Science Tripos Part II
Lent Term 2004

Simone Teufel

Natural Language and Information Processing (NLIP) Group

sht25@cl.cam.ac.uk

Information access in an ideal world

Question: "What was the historical development of Boolean algebra and set theory?"

Answer: "In 1854 George Boole published a seminal work *An investigation into the Laws of Thought, on Which are founded the Mathematical Theories of Logic and Probabilities.*"...

The user's information need is ideally met:

- Right type of response for information need
- Expected amount of information
- In perfect English, natural interaction
- And the information is of course correct!

Instead: Formulate question as a query containing terms, and review documents which the system returns

Q_1: set* bool* algebra

IR System manipulates terms:

Q_1': set* .467 boole* .751 algebra .091

via

$$weight_{w,q,D} = tf_{w,q} \cdot idf_{w,D}$$

Answers: documents, in order of estimated relevance 4

A':

[DOC 1 — 1.309]

... Boole's algebra, (that is in fact a set of algebra structures) was introduced in 1847 so as to propose an algebraic formulation of logical proposals. ...

[DOC2 — 1.286]

Answer your questions about algebra and other areas of mathematics by setting up your IR system – You can use Boolean queries.

[DOC3 — 1.211]

This was a breakthrough for mathematics and Boole was the first to prove that logic is part of mathematics and not of philosophy as was commonly accepted by scientists of this era.

In what ways can a document be relevant to a query?

- Answer precise question precisely
- Partially answer question
- Suggest a source for more information
- Give background information
- Remind the user of other knowledge
- Be objectively relevant, although user knows the document already

Types of information needs

- Precise information-seeking search $\rightarrow$ don't care where the information comes from, expect at least one document answering it
- Known-item search: Know that a certain item is there, want to refind it $\rightarrow$ want to find exactly that item
- Open-ended search ("topic search"): $\rightarrow$ do not know if a document exists; potentially, many exist

- Information scarcity problem (or needle-in-haystack problem): hard to find rare information
  - Lord Byron's first words? 3 years old? Long sentence to the nurse in perfect English?
- Information abundance problem (for more clear-cut information needs): redundancy of obvious information
  - What is toxoplasmosis?

A known-item search, in old-style library catalogue 8

"Boole's book"

- If I know title, author, year ("An investigation of the laws of thought, on which are founded the mathematical theories of logic and probabilities, George Boole, 1854")

  $\rightarrow$ Boole, G ... 1854 $\rightarrow$ Location: Betty & Gordon Moore Library Classmark: QA9 .B65
- "Bool? Boole? on algebra??, 19th Century" $\rightarrow$ L paper catalogue

  as above
- "on algebra, 19th century, called '*Laws of Thought*'" $\rightarrow$ L subject index

  Algebra 347.9 $\rightarrow$ Shelves

“Boole’s book”

Search	Query	Results
(full text)	“laws thought”	→ 10000 entries (truncated)
(title)	“laws thought”	→ 0 entries
(title)	“laws of thought”	→ 2 entries
(title)	“law of thought”	→ 0 entries
(title)	“algebra”	→ 623 entries
(full text)	“algebra”	→ 2474 entries
(Boolean)	“logic AND boole”	→ 9 entries

A known-item search, on Google

Google

- “law” → Lawyers, legal services, law schools
- “laws” → Lists of laws, public and private laws, four spiritual laws
  → law and laws are different search terms on Google!
- Boole does not show up in the first 10 pages

- "law thought"
  - "Savannah NOW: Local News - Mother-in-law thought mechanic was a ..."

→ Unexpected other meanings of terms (here: idiomatic)

→ Grammatical function of term ignored (verb instead of noun)

→ Treatment of dash

Rest: divorce, law as thought control, law and social thought, Scottish thought...

- "thought law"
  - "And you thought law enforcement was boring?"

→ Order of terms matters in Google (but capitalisation does not)

A better Google search: (develop|history|historical) set theory (boole|boolean) algebra

STEP II: Develop Course Objectives and Outcomes
STEP II: Develop Course Objectives and Outcomes COURSE ... partially ordered sets, lattices, Boolean algebras, semigroups ... and predicate logic 2. Set theory and its ...

06: Order, lattices, ordered algebraic structures
... History. ... especially on infinite sets is the study of Ordinals in Set Theory; ...

03E: Set theory
... do no better describing the history of Set ... propositions; 03E30: Axiomatics of classical set theory and its ... Other aspects of forcing and Boolean-valued models; ... Description: From Dave Rusin's "Known Math" collection.

Introduction to Algebra: History
... The next major development in the history of algebra ... Boole's original notation is no longer used, and ... uses the symbols of either set theory, or propositional ...

HiLight ... to a particular epoch in human history, that of ... Every historical position to achieve is like a ... Algebra Classical propositional logic and set theory are often ...

Selected course history
... Rosen). Basic set theory, discrete probability, combinatorics, Boolean algebra, graph theory. Fundamentals of Dynamical Systems. Elementary

MA003 MATHEMATICS FOR COMPUTING
... develop and use the concepts presented in the lectures. In particular emphasis will be put on parallel or similar systems such as set theory & Boolean algebra ...

Boolean algebra
... a new method of diagramming Boole's notation; this was ... When used in set theory, Boolean notation can demonstrate the ... indicating what is in each set alone, what ...

Graduate Courses
... Combinations, logic set theory, Boolean algebra, relations and functions, graph ... The historical evolution of non-Euclidean geometries ... History of Mathematics. ...

Barnes & Noble.com - Ones and Zeros: Understanding Boolean ...
... features include: a history of mathematical logic, an ... Electronic digital computers, Set theory, Design. Logic, Symbolic and mathematical, Circuits, Algebra, Boolean. ...

Somewhere in document 4:

> ... Boolean algebra was formulated by the English mathematician George Boole in 1847 ...

Somewhere in document 8:

> Boolean algebra, an abstract mathematical system primarily used in computer science and in expressing the relationships between sets (groups of objects or concepts). The notational system was developed by the English mathematician George Boole c.1850 to permit an algebraic manipulation of logical statements.

$\rightarrow$ Searching requires knowledge about underlying model to be effective

AltaVista

law*

- law
- laws
- lawyers
- lawn
- Lawndale, CA

- Which text representation is indexed? Abstract, title, full text? Can I specify where I am searching?
- Is the output ranked?
- If more than one search term is used, is each term guaranteed to be there?
- Can I exclude a term? Can I make a search term compulsary?
- Are my search terms stemmed?
- Are there wildcards?
- Does capitalisation matter?
- If more than one search term is present, is there an implicit “AND”? Which other boolean operators are available?

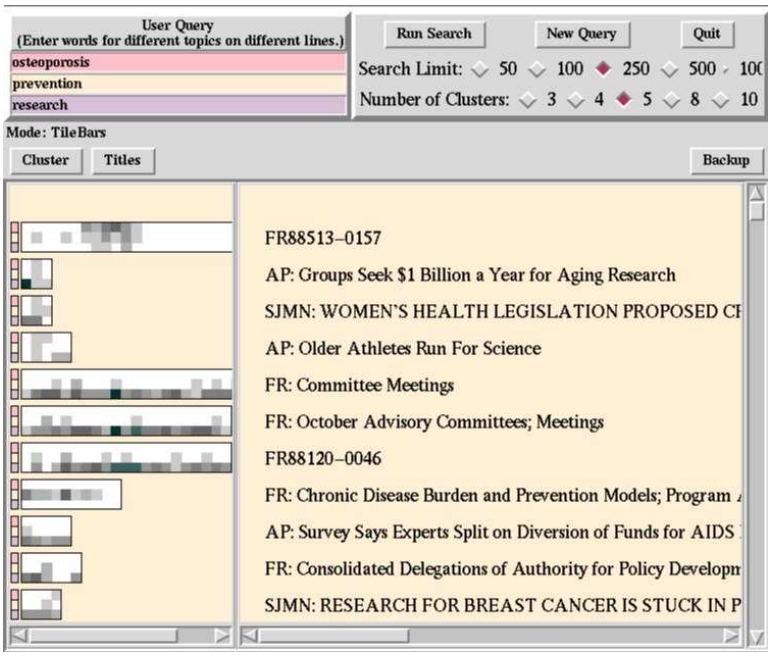
Factors in searching, continued

- If there are more than one search term, does the order matter?
- Can I use proximity? (Altavista’s NEAR: within 10 words of each other)
- Can I use phrases? (“laws of thought”)

Choosing search terms

- More specific terms are generally better
- Some terms mean other things too (think of different word senses, use exclusion)
- Other terms can mean the same thing (think of synonyms)
- Interference from WWW-specific key terms (“schema”, “domain”)

TileBars: Show distribution of query terms (e.g. “osteoperosis” “prevention” “research”) in segments of the document



Asking Questions

“What did George Bool invent?”

“And when?”

Solution One:

Ask google for the following exact strings:

```
"Boole invented"
"Boole developed"
"Boole invents"
"invented by Boole"
"invented by George Boole"
"developed by Boole"
"Boole discovered"
```

George Boole invented a mathematical tool for future computer-builders–an ”algebra of logic” that was used nearly a a hundred years later to link the process of human reason to the operations of machines.

George Boole invented a system for symbolic and logical reasoning, called Boolean Algebra, which became the basic design tool for computer design.

George Boole invented a bramch of mathematic colled Coolean Algebra which has been applied to the development of logic and electrical relays.

1854 AD, Boole invented Boolean algebra.

George Boole invented Boolean logic, which birthed discrete mathematics, which enabled the development of the transistor and ultimately the home computer, so essentially this one man is the reason our country is so fat, detached, sequestered, posh, and lazy and the reason we are all headed for a sublime Hell on Earth not unlike the environment depicted in the video ”Sober” by Tool.

Around the 1850s, the British mathematician George Boole invented a new form of mathematics, in which he represented logical expressions in a mathematical form now known as Boolean Algebra.

Boole invented ”Boolean algebra” (switching theory)

Boole invented the truth table to test the truth and validity of compound propositions.

Boole invented the first practical system of logic in algebraic form

While working here in UCC in 1854, Boole invented Boolean Algebra which has become the cornerstone of modern electronics and information technology.

George Boole invented propositional logic (1847).

In 1854, Boole invented Boolean algebra. George Boole invented the branch of mathematics known as Boolean algebra. In the 19th century George Boole invented Boolean algebra as a theoretical study.

Later in the year 1850, Charles Boole invented binary codes, which uses only numbers 0 and 1.

In the 1850s, George Boole invented a mathematical system of symbolic logic that would later become the basis for modern computer design.

The English mathematician, George Boole, developed an algebra of logic, which has become the basis for computer database searches.

Boole developed an algebraic calculus to interpret whether composite assertions were true or false

Initially a schoolteacher in Lincolnshire and Yorkshire, England, Boole developed his ideas about symbolic logic without the benefit of a formal university training.

Boole developed ”The Mathematical Analysis of Logic.”

Boole developed the meaning of the logical operators.

By considering assertions to by true or false, Boole developed an algebraic calculus to interpret whether composite assertions were true or false in terms of how they composition was formed.

The first major advance came when George Boole developed an algebra of logic.

The Algebra of Logic was originally developed by Boole

Boolean algebra was developed by Boole.

The modern concept of abstract algebra was developed by Boole In the 1930’s Claude Shannon, an American mathematician who later worked for Bell Labs, noted that the algebra developed by Boole was the appropriate

The idea of a completely formalistic logic, however, was developed by Boole in the 19th century.

an internationally recognized system of logic invented by Boole

Algebra invented by Boole allows easy manipulation of symbols

Probabilistic logic, invented by Boole, is a technique for drawing inferences from uncertain propositions for which there are no independence assumptions.

This page defines ’boolean,’ pertaining to the logical operations described in the algebra invented by George Boole

That little bit of computer logic was actually invented by George Boole, a British mathematician and logistician.

Shannon explained how the algebra invented by George Boole in the mid-1800s could be used to . . .

Boolean circuits are certain electric circuits that were invented by George Boole.

So are the bit and bytes that obey complex laws first invented by George Boole a hundred or so years ago.

Search engines use Boolean logic, which is a system of logic invented by George Boole, a nineteenth century mathematician.

Explain how Boole applied his new algebra to logic, and explain (according to Davis) how it was that Boole discovered the importance/validity of using the sets 1 and 0 by proceeding in accord with the Aristotelian ”principle of contradiction” Bayes's work.

Boole discovered analogy of algebraic symbols and those of logic.

In 1854 George Boole ”discovered pure mathematics”(Bertrand Russell's expression) this equivalence

How to recognise that... 22

- ... the following are not answers:
  - What Boole discovered in that meadow and worked out on paper two decades later was destined to become the mathematical linchpin that coupled the logical abstractions of software with the physical operations of electronic machines.
  - Lots of Rules and EXECs, developed by Boole SSEs and customers.
- ... the following might be answers, but more work is needed:
  - Boole invented this method back in the 18th century, so that human thought could be strictly analysed and evaluated.
  - It was invented by George Boole, a British mathematician in the 1840's.
  - Leibniz ... invented binary numbers in this case, but he didn't even invent propositional calculus; that was invented by Boole one hundred and fifty years later.
- ... the last sentence cancels our sought-for fact:
  - Many collegues in Algebra and Logic think that Boole developed either Boolean Algebra, or Boolean Rings. He did neither.

We implicitly used information extraction in the QA-“shortcut” solution.

But the string-based approach does not generalise to similar mentions, eg. “Boole was the inventor of”

How does information extraction work?

- Named Entity Recognition:
  Find entities of a certain semantic type in unrestricted text

  HERE: Find names, find dates, in all possible formulations, with robustness to typos and formatting
- Coreference Resolution:
  Decide which strings refer to the same entities
- Template Recognition:
  Find predefined relations in unrestricted text, for example inventors/invention patterns

IE templates

The INVENTOR relation/template has pre-determined slots and relationship between slots

- “To accomplish this, we’ll first learn about the concept of Boolean algebra - a system of logic designed by George Boole.”

  INVENTOR: George Boole
  INVENTED: Boolean algebra
- “Peirce developed what amounts to a semantics for three-valued logic.”

  INVENTOR: Peirce
  INVENTED: a semantics for three-valued logic

Necessary if no handy meaning–surface mapping is available and/or if there isn't enough data:

- Need to understand something about the question:
  - Grammatical function and thematic roles – who does what to whom
  - What is the expected answer type?
- Need to understand something about the document:
  - Where is the expected answer type in the text?
  - Not all information is locally available in the answer sentence (e.g. anaphora, list discourse markers. . .)
- Need to produce an answer:
  - Answers are phrases, sentences, paragraphs, 50Byte strings
  - "Information packaging", reversal of non-local effects

Summarisation

- The holy grail of NLP
- Methods working today are either very simple or very complicated
- The simple ones are robust but have many other disadvantages
  - Textual problems for all effects above the sentence-level
  - No guarantee of truth preservation
- The complicated ones are not robust
- Two recent tasks:
  - Multi-document summarisation
  - Incremental-time-line summarisation
- Evaluation is a major problem

- Each sentence is represented by a set of 'importance indicators' (features)
- These are combined in such a way to rank the sentences
- The N highest-ranking sentences are extracted and constitute the summary (here: black sentences)

		Importance indicators			
1	Algebra provides a generalization of arithmetic by using symbols, usually letters, to represent numbers.	0	1	3	1
2	For example, it is obviously true that 2 + 3 = 3 + 2	0	0	2	1
	...				
14	In about 1100, the Persian mathematician Omar Khayyam wrote a treatise on algebra based on Euclid's methods.	0	0	1	0
	...				
26	Boolean algebra is the algebra of sets and of logic.	1	0	1	1
27	It uses symbols to represent logical statements instead of words.	1	0	1	1
28	Boolean algebra was formulated by the English mathematician George Boole in 1847.	1	1	2	1

		Importance indicators			
29	Logic had previously been largely the province of philosophers, but in his book, The Mathematical Analysis of Logic, Boole reduced the whole of classical, Aristotelian logic to a set of algebraic equations.	0	0	0	1
30	Boole's original notation is no longer used, and modern Boolean algebra now uses the symbols of either set theory, or propositional calculus.	0	0	0	1
31	Boolean algebra is an uninterpreted system - it consists of rules for manipulating symbols, but does not specify how the symbols should be interpreted.	0	0	0	1
32	The symbols can be taken to represent sets and their relationships, in which case we obtain a Boolean algebra of sets.	0	1	3	0
33	Alternatively, the symbols can be interpreted in terms of logical propositions, or statements, their connectives, and their truth values.	0	0	0	1
34	This means that Boolean algebra has exactly the same structure as propositional calculus.	0	0	0	1
35	The most important application of Boolean algebra is in digital computing.	1	0	2	1
36	Computer chips are made up of transistors arranged in logic gates.	0	0	0	1
37	Each gate performs a simple logical operation.	0	0	0	1
38	For example, an AND gate produces a high voltage electrical pulse at the output r if and only if a high voltage pulse is received at both inputs p, q.	0	0	0	1
39	The computer processes the logical propositions in its program by processing electrical pulses - in the case of the AND gate, the proposition represented is p q r.	0	0	0	1
40	A high pulse is equivalent to a truth value of "true" or binary digit 1, while a low pulse is equivalent to a truth value of "false", or binary digit 0.	0	0	0	1
41	The design of a particular circuit or microchip is based on a set of logical statements.	0	0	0	1
42	These statements can be translated into the symbols of Boolean algebra.	0	0	0	1
43	The algebraic statements can then be simplified according to the rules of the algebra, and translated into a simpler circuit design.	0	3	1	1
44	An algebraic equation shows the relationship between two or more variables.	0	0	0	1
45	The equation below states that the area (a) of a circle equals p (pi, a constant) multiplied by the radius squared (r 2).	0	0	0	1

Doc1
Born in Lincoln, England on November 2, 1815, George Boole was the son of a poor shoemaker. As a child, Boole was educated at a National Society primary school. He received very little formal education, but was determined to become self-educated.

Doc2
George was born in 1815 at 34 Silver Street, which is now occupied by Langleys, the Solicitors. He was the eldest son of John Boole, a shoemaker.

Doc3
Born in the English industrial town of Lincoln, Boole was lucky enough to have a father who passed along his own love of math.

Doc4
Boole was born of humble parents in 1815, the same year as the battle of Waterloo. It is doubtful he received early schooling in mathematics beyond that required for the most basic commerce. Unsatisfied with mathematics texts of the time, he set about reading the great masters, Gauss, Laplace, Liebnitz, and others.

Re-generation after clustering

Cluster 1:
Born in Lincoln, England on November 2, 1815, George Boole was the son of a poor shoemaker. George was born in 1815 at 34 Silver Street, which is now occupied by Langleys, the Solicitors.
→ S1: “George Boole was born in 1851”.

Cluster 2:
Born in Lincoln, England on November 2, 1815, George Boole was the son of a poor shoemaker. Born in the English industrial town of Lincoln, Boole was lucky enough to have a father who passed along his own love of math.
→ S1’: “George Boole was born in 1851 in Lincoln, England.”. (Correction; aggregation)

Cluster 3:
Boole was born of humble parents in 1815, the same year as the battle of Waterloo. He was the eldest son of John Boole, a shoemaker. Born in Lincoln, England on November 2, 1815, George Boole was the son of a poor shoemaker.
→ S2: “He was the son of a shoemaker”.

Cluster 4:
As a child, Boole was educated at a National Society primary school. He received very little formal education, but was determined to become self-educated.
→ S3: “He received little schooling as a child.”

- Why searching is difficult: mapping from natural language to the underlying search model
- Query constructs: Search terms and how to combine them
- Retrieval models: Finding the best answer document
- Beyond Information Retrieval: Information extraction
  - Information is known beforehand
  - Information need can be expressed by templates
- Question answering: domain-independent
- Summarisation: additional problem of text production

This course: topics

IR
- Lecture 2: Information retrieval models
- Lecture 3: IR evaluation methodology
- Lecture 4: Search engines and linkage algorithms

IE
- Lecture 5: Information extraction
- Lecture 6: Bootstrapping in information extraction

Lecture 7: Question answering

Lecture 8: Summarisation; outlook