The World-Wide Web: Quagmire or Goldmine?

Oren Etzioni
[Comm. of the ACM, Nov 1996]

Presentation Credits:
Shabnam Sobti

Agenda

- Prelude: The Internet Story
- Article Review: Mirrored Reflections
- An Individual View: Multiple Facets
- Intermediate Development and Future Scope: Roadmapping the Third Front
- Sum - up: Cadenza
The Internet Story

“The time has come, the Walrus said, 
To talk of many things. 
Of shoes and ships and sealing wax—
Of cabbages and kings.”

→ Lewis Carroll

“Through the Looking Glass”

- Expansion: necessity
- ? Moral? “…… will anybody really live happily ever after?”

Mirrored Reflections: The Article’s Perspective

“Fine goals matter less than the right strategy.”

→ Stanley Hoffman

Information Food Chain

Carnivores

Herbivores

Grazers

SoftBots: AOH1; MetaCrawler - Personal Assistants

Indices, Directories - Alta Vista, Yahoo, Mass Services

WWW Pages and Hyperlinks, Mass Services
Mirrored Reflections: Introduction / 1

- **Basic Question:** “Is Information on the Web sufficiently well-structured to facilitate effective Web Mining?”

- **Issues:**
  - Buried information in Data Mining.
  - No machine readable semantics.

- **Possible Solutions:**
  - Transform Web into massive, layered database.
  - Hand-code specific wrappers.

Mirrored Reflections: Introduction / 2

- **Alternative to Structured Web Hypothesis**

- **Structure:**
  - Linguistic & typographic conventions.
  - HTML annotations
  - Classes of semi-structured documents
  - Web indices, directories, etc.

- **Organization of Web Mining:**
  - Resource Discovery
  - Information Extraction
  - Generalization
2 classes of web resources:
- Documents
- Services
- Focus: automatic creation of searchable web indices.
- Egs:
  - Alta Vista
  - MetaCrawler

AltaVista - Characteristics:-
- Massive memory and network bandwidth requirements
- Cost of resources is curtailed
- Independent queries; no customization
- Homogenised, LCD service

AltaVista - Attributes:-
- Scan documents – store index of words
- Ask for indexed documents using keywords

AltaVista - Drawbacks:-
- Repetition of queries
- Irrelevant responses
Mirrored Reflections: Resource Discovery / 3

- MetaCrawler – Characteristics:
  - Interface and query language
  - Collation and pruning
  - Local phrase search
  - Web services and interfaces decoupling
  - Meta interface - benefits:
    - Modest access
    - Customization
    - No need to downsize ‘smartness’

- MetaCrawler – Evolution Scope:
  - Document clustering
  - Mixed-initiative dialog

Mirrored Reflections: Resource Discovery / 4

- Future resource discovery systems:
  - Automatic text categorization
  - Automatic construction of web directories
  - Filter query results to searchable indices
Mirrored Reflections: Information Extraction / 1

- **Challenge:** automatic extraction of information from a discovered source.
- **Current status:**
  - Identify fixed set of resources
  - Hand coded wrappers for parsing
- **Need:** dynamic extraction from unfamiliar sources

Mirrored Reflections: Information Extraction / 2

- **Harvest:**
  - Models of semi-structured documents.
  - No discovery of new documents.
  - No learning of new models of document structure.
  - Easy handling of familiar types.
Mirrored Reflections:
Information Extraction / 3

- FAQ-Finder:
  - More potential of returning higher quality information.
    - Semi-structured files
    - Smaller number of files
- Limitations [both]:
  - Focus on documents; ignore services
  - Pre-specified description

Mirrored Reflections:
Information Extraction / 4

- Internet Softbots:
  - Test queries + domain specific knowledge = auto learning of web service descriptions
  - Egs:
    - ILA
    - Shopbot
ILA [Internet Learning Agent]:
- Automatic models of declarative learning
- Query unfamiliar resources against known objects
- Competing hypothesis
- Requirements suited for agents having no formal description
- Drawbacks:
  - Category mismatch
  - Token mismatch
  - Conjunctive bias

ShopBot - Characteristics:
- Extraction using minimal knowledge
- Ambitious task: learn by querying and response analysis
- Operates in 2 phases:
  - Learning
  - Comparison shopping
- Scalable and robust
- O/P structure and prototypical queries
Mirrored Reflections:
Information Extraction / 6

ShopBot:
- Limitations:
  - Doesn’t understand meaning of description
  - Strong bias
  - Extraction of individual parts
- Current work: autonomous discovery of vendor home pages

Mirrored Reflections:
Generalizations / 1

- Most m/c learning systems learn about user interests instead of the Web
- Obstacle: Labelling
- Inputs labelled as (+ve) /(-ve) sample
- Techniques so far:
  - Uncertainty Sampling
  - Clustering
- Solution approach: Web is much more than collection of linked documents
Mirrored Reflections: Generalizations / 2

Ahoy! :-
- Harness the power of users
- Queries MetaCrawler and filters its output
- Heuristic filtering algorithm
- Solves labelling problem using user feedback
- Advantages:
  - Rapid collection of learning data
  - Architecture not restricted to particular learning domain

Mirrored Reflections: Conclusions

- Very brief, selective survey
- Intuitive hypothesis proved by citing examples
- Enormous Web potential
- Although the Web is less structured than we think – its not as random as we fear
Multiple Facets: An Individual View / 1

"Technology is the Extension of the Central Nervous System."

Marshall Mc.Luhan

Van nevar Bush

DATA

INFORMATION

KNOWLEDGE

Correctly Structured

Correctly Applied

Multiple Facets: An Individual View / 2

Three sides to every story – yours; theirs and the right one.

Effectively an article

Written in 1996

Therefore, the hypothesis question is a paradox in itself !!
Roadmapping the IIIrd Front:
Intermediate Development and FutureScope / 1

The only reason for progress is the desire of any organism to live beyond its means.

- **SoftBots:**
  - Expectations from Personal Assistants: Robustness; Speed; Added Value
  - Intelligent agents that use s/w tools and services on user’s behalf
  - Criticality of AI: Patrick Winston’s “Raisin-Bread” model
  - Advantages:
    - Cost, effort, expertise needed is low
    - S/w environments circumvent certain problems
    - Simulated physical worlds take a long time to perfect
    - Eg: Rodney

Roadmapping the IIIrd Front:
Intermediate Development and FutureScope / 2

- **WebML:**
  - Approach: tailored wrappers that map document features
  - Resource as well as knowledge discovery
  - Interactive querying
  - Assets:
    - Relatively high level concept mapping
    - Unique knowledge discovery power
    - Advantage of MLDB model
  - Takes advantage of MLDB model
  - Propagation algorithm
Issues with AI scalability:
- Discovery, Extraction, Translation, Evaluation

Extraction and Translation:
- ShopBot, ILA, Levy & Ordille’s model and Semint

Discovery:
- Agents monitor “what’s new”
- Agents investigate user sent URL(s)
- Co-operation with existing tools

Evaluation:
- Source testing
- Comparison and validation
- Motro – Rakov model

AUTO-FAQ:
- Presumption: cost of acquisition of knowledge too high
- Cyber leveraging – CYLINA

Features:
- Shallow Language understanding
- Population Leveraging for information acquisition and adaptive information management
- Q – A orientation
- Cyber leveraging through: identification of gaps in the infobase, adding knowledge to plug them and gather feedback.
Cadenza
A Summary

- Structural attributes of the web are subjective
- Point of views of naïve and experienced users, developers
- Data on the web always liable to better reforms – alleviate randomness
- ? – whether the quagmire can be turned into a goldmine.

Cadenza
Afterthoughts

- Cannot start with the presumption that the said hypothesis is incorrect.
- Research and development, optimality, scalability, etc. are extensible concepts.
- Therefore, the discussion must not focus on the (in)correctness of the hypothesis – but brainstorm alternative areas of further research.
THANK - YOU