 Identify the Intelligence Values of Web Resources Based on Knowledge Object Grids

Zhang Zhixiong, Zou Yiming, Liu Jianhua, Xie Jing, Qian Li

National Science Library, Chinese Academy of Science

Sept.5, 2012
Outline

1. Background Information
2. Intelligence Value
3. New Ideas based on Object Grid
4. Analysis and Discussions
Outline

1. Background information
2. Intelligence Value
3. New Ideas based on Object Grid
4. Analysis and Discussions
1. Background

- National Science Library, CAS
  - Main library in Beijing
  - 3 Branch libraries in
    - LanZhou
    - ChenDu
    - WuHan
1. Background

- National Science Library, CAS
  - total 470 staffs, about 90 of them work for information analysis services
    - monitor scientific research activities
    - identify the research trends
    - detect the progress of the research areas
    - support decision maker to develop research policy for China

- Every year, they publish
  - about 60 issues of Special Edition on important research & innovation activities
  - about 270 issues of Monitoring Express on the progress of research in 13 categories
1. Background

- So...it is very important to monitor the research & innovation activities published from key institutes website, such as:
  - national science & innovation administrative offices
    - OSTP (the Office of Science and Technology Policy)
  - research councils
    - Research Councils UK
  - government department
    - Department of Energy
  - funding agencies,
    - NSF, European Science Foundation
  - leading research organization
    - NAS, the Academy Sciences for developing world, NIH
  - innovative research and analysis institutes
    - Brookings Institution
1. Background

- Monitoring system built, to help information analyst team to:
  - obtain timely the new web pages from the key research institutes
  - judge the intelligence value of these web pages
  - disclose the main content of the new web pages
  - calculate the topics and objects (the person, the institutes, the projects...) embedded in these resources
  - visualize the patterns and trends base on the calculations
  - ......
1. Background

- Up to now, several monitoring services have been set up for some research areas
1. Background

- "Object-based Computing" methodology in Research Profiling

  - Object
    - named entities embedded in web pages, such as
      - science strategies and policies, "A Strategy of American Innovation"
      - key initiatives & research programs, "Framework 8"
      - key institutes, "OSTP", "Department of Energy"
      - ...

  - Object-based Computing
    - extract the objects from the web page and do calculation of those objects to disclose the pattern and trends of changing situation
From Mohawks to Making, New Steps to Mobilize the Science Talent in Federal Agencies

Subscribe

From Mohawks to Making, New Steps to Mobilize the Science Talent in Federal Agencies

Posted by Kumar Garg and Phil Larson on August 21, 2012 at 05:26 PM EDT

Bobak Ferdowsi-or NASA’s “Mohawk Guy”. (Photo courtesy NASA)

President Obama strongly believes that inspiring boys and girls to excel in science, technology, engineering and math (STEM) education is critical to our Nation’s future. Just last week, the President called the Mars Science Lab Team and a “special Mohawk guy” and pointed out how their work on Curiosity was inspiring the next generation.

As the President said, “My Administration has put a ‘hands-on’ on improving science and technology, engineering and math education. And this is the kind of thing that inspires kids across the country. They’re telling their moms and dads they want to be part of a Mars mission—maybe even the first person to walk on Mars. And that kind of inspiration is the byproduct of work of the sort that you guys have done.”

The Curiosity team is emblematic of the President’s call that the Federal science and technology talent can be in the vanguard of an effort to improve STEM education.

That’s why President Obama has called upon the 200,000 Federal employees working in STEM fields to bring their passion and expertise to their communities and schools in support of STEM education, and help “stoke that some curiosity in students which had perhaps led them to pursue a career in science.” As the President has said, there are so many creative ways to engage young people in STEM fields—everything from science festivals, robotics competitions, Maker Faires, mentoring opportunities and more.

To build on the President’s call to action, Office of Personnel Management (OPM) Director John Berry has signed a memo that further encourages the talented men and women serving in the Federal STEM workforce to volunteer their time and expertise towards improving STEM education.
1. Background

monitoring all the targeted institutes timely

<table>
<thead>
<tr>
<th>Rank</th>
<th>Institute Name</th>
<th>Country</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>France Space Research Center (CNES, 空间研究机构, 法国)</td>
<td>France</td>
<td>Founded in 1961, the Centre National d'Etudes Spatiales (CNES) is the government agency responsible for shaping and implementing France's space policy in Europe.</td>
</tr>
<tr>
<td>2.</td>
<td>DFG, 科学联合会, 德国</td>
<td>Germany</td>
<td>DFG must remain abreast of current developments in higher education.</td>
</tr>
<tr>
<td>3.</td>
<td>Max Planck Society, 德国科学院, 德国</td>
<td>Germany</td>
<td>The mission of the Office of Science and Technology Policy is threefold, first to provide the President and his senior staff with accurate, relevant, and timely scientific and technical advice on a...</td>
</tr>
<tr>
<td>4.</td>
<td>Research Councils UK, 研究委员会, 英国</td>
<td>UK</td>
<td>RCUK was a strategic partnership between the seven UK Research Councils. RCUK was...</td>
</tr>
</tbody>
</table>
1. Background

- monitoring all the targeted institutes timely
- keeping trace of recent development of targeted research activities
1. Background

monitoring all the targeted institutes timely

keeping trace of recent development of targeted research activities

identifying intelligence value automatically
1. Background
1. Background

- monitoring the targeted institutes timely
- keeping trace of recent development of targeted research activities
- Identifying intelligence Value automatically
- automatically analyzing the web pages and rich content documents
- automatically identifying key research activities and key research topics
1. Background

- Monitoring all the targeted institutes timely
- Keeping trace of recent development of targeted research activities
- Identifying intelligence Value automatically
- Automatically analyzing the web pages and rich content documents
- Automatically identifying key research activities and key research topics
- Automatically identify hot research activities and hot topics
1. Background

monitoring all the targeted institutes timely

keeping trace of recent development of targeted research activities

Identifying intelligence value automatically

automatically analyzing the web pages and rich content documents

automatically identifying key research activities and key research topics

automatically identify hot research activities and hot topics

provide edit tools for creating Monitoring Express
Outline

1. Background
2. Intelligence Value
3. New Ideas based on Object Grid
4. Experimental Results
5. Analysis and Discussions
2. Intelligence Value

importance of each web resource to information analyst:

Should be:

Five Star?
Four Star?
Two Star?
Or Less?
2. Intelligence value

5-Dimensional Intelligence Value Indicator System

Intelligence Value Indicators Systems

Objects Dimensional
- Research object
  - Obama, Innovation, Policies or a visitor speech

Source Dimensional
- Research Policies from OSTP website
- NAP reports from NAS

Content Type Dimensional
- Strategy Plan, R & D Budget, Report, or just a news

5-Dimensional Indicator System

Domain Dimensional
- political, economical, diplomatic, or S & T related
- Energy, Resource and Environment, Space....
2. Intelligence value

- The good sides of current intelligence value indicator system
  - simple
  - easy to understand, easy to implementation

- But
  - hard human workforce
    - We need a training set firstly.
    - set the values for the indicators early...
  - can not automatically adjust the value of the indicator with the thing go on
    - Big Data
2. Intelligence value

- Can we find a way that can identify the Intelligence value of web resources more automatically?
  - Automatically set the importance value for the indicators?
  - Automatically judge the importance of the web resources by more automatically identify the key objects in the web resources.
- So we come to the ideas of objects grid
Outline

- Background Information
- Intelligence Value
- New Ideas Based on Object Grid
- Analysis and Discussions
3. New Ideas based on Object Grid

- **Object Grid come from Entity Grid**
  - proposed by Barzilay and Lapata((2008, representing and measuring local coherence)
  - An entity-grid, representing contents contained in text.
  - This 2-D grid (or matrix) shows presence or absence of entities for each sentence and whether the entity is a subject (s), object (o), or others (x).
3. Idea of Knowledge Object Grid

- Entity Grid

Entity

subject (s), object (o), or neither (x)

Text

Entity Grid

Entity Grid: Proposed by Barzilay and Lapata
3. Idea of Knowledge Object Grid

- **Entity Grid**
  - is a good way to
    - disclose the important entities in the text.
    - segment the text into connected blocks
  - but, it has some limitations to be extended
3. Idea of Knowledge Object Grid

- The limitations of Entity Grid
  - Use head nouns in stead of formal named entities and terms
    - single words can't present definite meanings and topics of one text explicitly
    - eg. the Justice Department vs Department
  - Care less about Co-reference. Many entities in fact are same, that could be used to connect text blocks.
    - eg. The President, President Obama, Barack Obama
  - Entity grid doesn't contained Predicate, which is a indictor the relationship of Subject and Object.
    - subject (s), object (o), or others (x)
3. Idea of Knowledge Object Grid

- **Object Grid**
  - Object grid extends entity grid in three aspects
    - Using Objects(named entities and terms) to replace head nouns.
    - Perform Co-reference to cluster objects semantically and using semantic chain to connect related objects.
    - Constructing SPO(subject–predicate–object) relations to represent relationships between the objects.
3. Idea of Knowledge Object Grid

- Object Grid
  - (1) Using Objects, instead of head nouns.
    - Defining eleven semantic classes for named entities, such as Person, Foundation, Conference and Project.
    - Usually, these research objects carry the core information of the web pages and are valuable for automatically extracting intelligence from web pages.

- "Department" VS "<Department of Energy  typeOf Key Government Department>"
3. Idea of Knowledge Object Grid

- Object Grid
  - (2) Perform Co-reference, create a semantic chain to connect related objects.
3. Idea of Knowledge Object Grid

- Object Grid
  - (3) Constructing SPO relations to represent relationships between the objects.
    - (Object A, Object B, Relationship Type)
    - Example:
3. Idea of Knowledge Object Grid

- **Object Grid**

  - Text
  - Knowledge Object Recognition
  - Coreference Disambiguation
  - Semantic Clustering
  - Syntactic Analysis
  - Knowledge Object Grid

  ![Diagram](image_url)

  By Ng and Cardie [1]
  - Stanford CoreNLP and Gate.
  - WordNet
  - Research Ontology
  - By Collins [2]

[1] Ng, C. Cardie. Improving machine learning approaches to coreference.
3. Idea of Knowledge Object Grid

Object Grid

1. [NASA's Voyager 1 spacecraft]$_s$ has entered a [new region]$_o$ between [our solar system]$_x$ and [interstellar space]$_x$.
2. [Data]$_s$ obtained from [Voyager]$_s$ over the last year reveal this [new region]$_s$ to be a kind of [cosmic purgatory]$_s$.
3. In [it]$_x$, [the wind of charged particles]$_s$ streaming out from [our sun]$_x$ has calmed, [our solar system's magnetic field]$_s$ has piled up, and [higher-energy particles]$_s$ from inside our [solar system]$_x$ appear to be leaking out into [interstellar space]$_x$.
4. "[Voyager]$_s$ tells us now that we're in [a stagnation region]$_o$ in [the outermost layer of the bubble]$_x$ around [our solar system]$_x$," said [Ed Stone]$_s$, [Voyager project scientist]$_x$ at [the California Institute of Technology]$_x$ in [Pasadena]$_x$.
5. "[Voyager]$_s$ is showing that what is outside is pushing back. We shouldn’t have long to wait to find out what [the space]$_x$ between [stars]$_x$ is really like."
6. Although [Voyager 1]$_s$ is about [11 billion miles (18 billion kilometers)]$_o$ from [the sun]$_x$, it is not yet in [interstellar space]$_x$.
7. In the latest data, [the direction of the magnetic field lines]$_s$ has not changed, indicating [Voyager]$_s$ is still within [the heliosphere]$_o$, [the bubble of charged particles the sun]$_s$ blows around itself.

Voyager Hits New Region at Solar System Edge  [OL].[2011-12-06].
http://www.space - ravel.com/reports/Voyager_Hits_New_Region_at_Solar_System_Edge_999.html
3. Idea of Knowledge Object Grid

- **Object Grid**

```
   1  S  O  X  X  -  -  -  -  -  -  -  -  -  -  -  -
   2  O  S  -  -  S  O  -  -  -  -  -  -  -  -  -  -
   3  -  X  X  X  -  -  S  S  -  -  -  -  -  -  -  -
   4  S  -  X  -  -  -  -  -  X  X  S  X  X  X  -  -
   5  S  -  -  X  -  -  -  -  -  -  -  -  -  -  -  -
   6  S  -  X  X  -  -  -  -  -  -  -  -  -  -  -  -
   7  S  -  -  -  -  -  S  -  S  -  -  -  -  -  -  O
```
3. Idea of Knowledge Object Grid

- Identifying the Important Knowledge Objects

![Diagram showing the process of identifying knowledge objects](Diagram.png)
3. Idea of Knowledge Object Grid

- Semantic Block segmentation
  - Text be split into a number of semantic blocks.

The goal of the Babel Program is to develop methods to build speech recognition technology for a much larger set of languages than has hitherto been addressed. The Program will require innovations in how to rapidly model a novel language with significantly less training data that are also much noisier and more heterogeneous than what has been used in the current state-of-the-art. Babel's technical measures of success are focused on how well the generated model works to support effective word-based search of noisy channel speech in the languages to be investigated. The new methods will be systematized so that they can be applied rapidly to a novel underserved language.
3. Idea of Knowledge Object Grid

- Identifying the local important objects of semantic block
  - Objects with larger span in the object grid are more important.
  - Objects with higher density are more important
  - Objects that bridge the important objects are more important
  - Objects co-occur with the important objects are more important
  - .......
3. Idea of Knowledge Object Grid

- Finding the global important objects from local important objects
- Calculate the intelligence values base on global important objects.
Outline

- 1. Background
- 2. Intelligence Value
- 3. New Ideas based on Object Grid
- 4. Analysis and Discussions
4. Analysis and Discussions

- We do think the new idea of Object Grid can improve the accuracy of Intelligence value.
- But, we still have not done enough
  - The experiment is still on the way...
  - We need to prove it.

- If Predicate has been added into Object Grid, the grid may be 3-Dimensional grids. It is really promising, but we need to do more to put it into good use.
Thanks!

Thank You for Your Attention!

谢谢!

zhangzhx@mail.las.ac.cn