

iSTAR and Text Mining

**Presentation to the iSTAR users' group
March 16, 2007**

Peter Chew (Sandia National Laboratories)



Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy under contract DE-AC04-94AL85000.





Content of presentation

- **Sandia's role: the NM Small Business Assistance Program**
- **What is text mining?**
- **How text mining has been used in other domains**
- **Applications of text mining to iSTAR data - ideas**
- **What are your ideas?**



What is text mining and why do it?

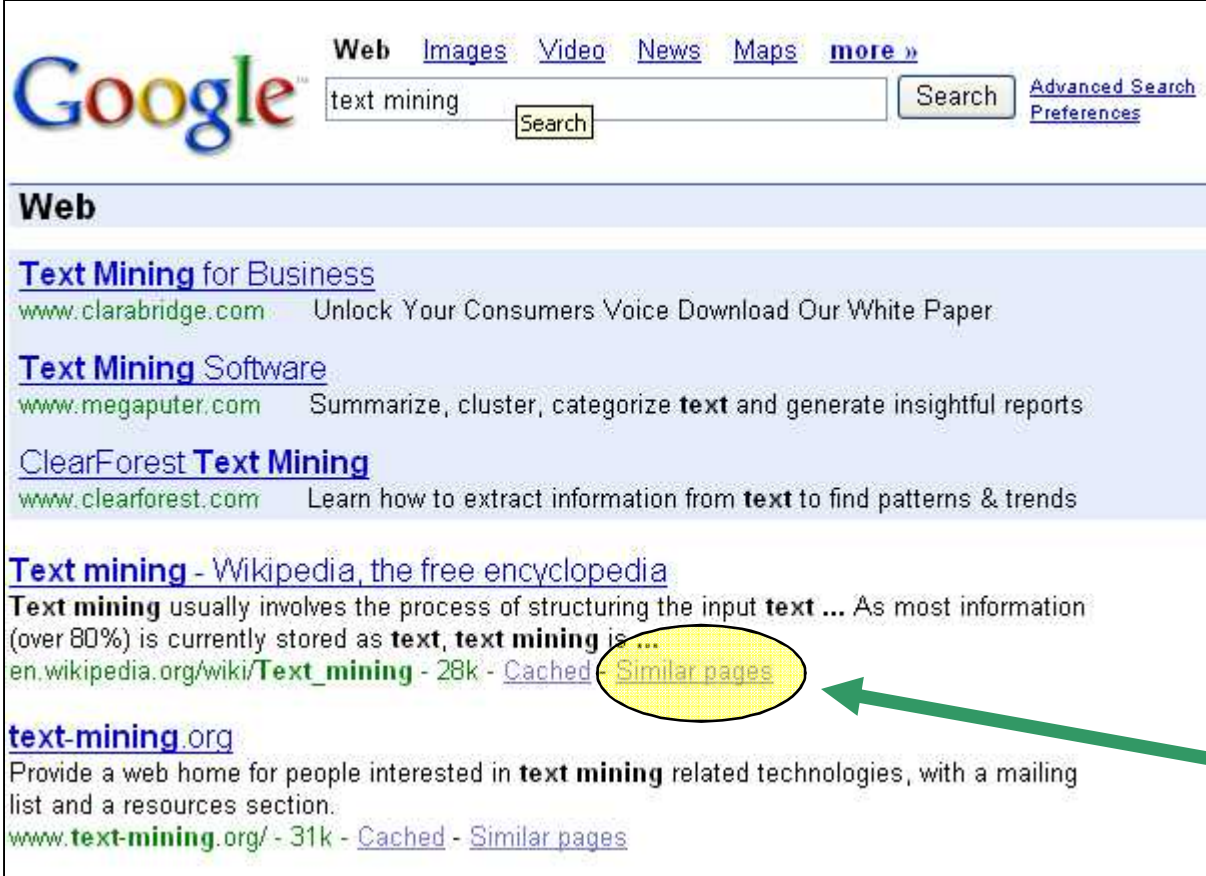
- ‘Extraction of interesting and non-trivial information and knowledge from unstructured text’
- 80% of data is unstructured text
- Text mining aims to extract the ‘concepts’ from text – it’s not just searching by keywords
- Finding the ‘needle in the haystack’
- Finding text which is similar to other text
- Understanding the big picture from large volumes of text



Finding the 'needle in the haystack'



Finding similar text



A screenshot of a Google search results page for the query "text mining". The page shows the Google logo, navigation links (Web, Images, Video, News, Maps, more »), and a search bar with the text "text mining". Below the search bar, there are several search results. The first three results are from commercial websites: "Text Mining for Business" from clarabridge.com, "Text Mining Software" from megaputer.com, and "ClearForest Text Mining" from clearforest.com. The fourth result is from Wikipedia, titled "Text mining - Wikipedia, the free encyclopedia". This result is highlighted with a yellow background. A green arrow points to the "Similar pages" link in the Wikipedia result. The Wikipedia result text includes: "Text mining usually involves the process of structuring the input text ... As most information (over 80%) is currently stored as text, text mining is ...". Below the Wikipedia result is another result for "text-mining.org".

Google

Web Images Video News Maps more »

text mining Search

Advanced Search Preferences

Web

[Text Mining for Business](#)
www.clarabridge.com Unlock Your Consumers Voice Download Our White Paper

[Text Mining Software](#)
www.megaputer.com Summarize, cluster, categorize **text** and generate insightful reports

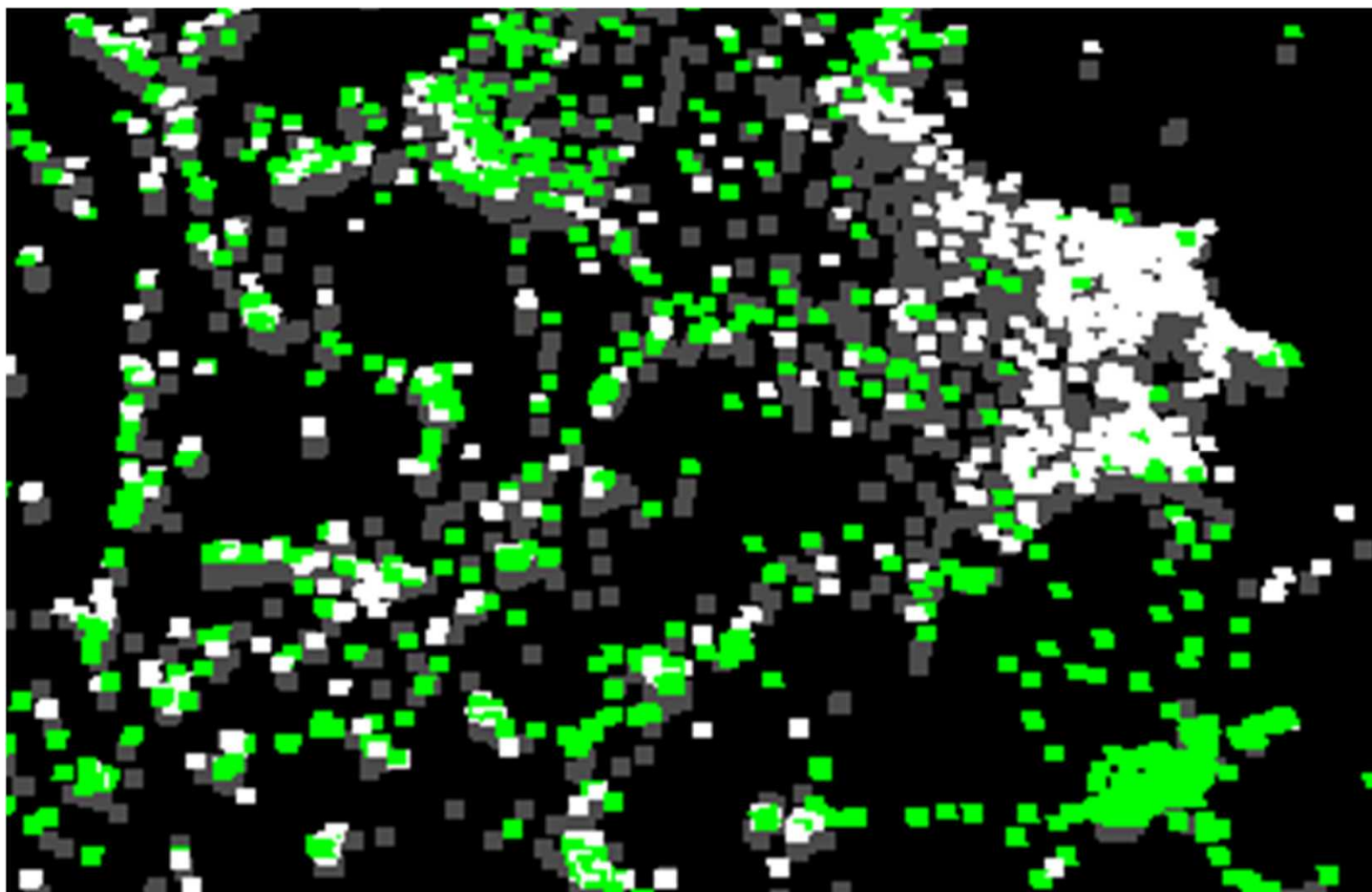
[ClearForest Text Mining](#)
www.clearforest.com Learn how to extract information from **text** to find patterns & trends

[Text mining - Wikipedia, the free encyclopedia](#)
Text mining usually involves the process of structuring the input **text** ... As most information (over 80%) is currently stored as **text**, **text mining** is ...
en.wikipedia.org/wiki/**Text_mining** - 28k - [Cached](#) - [Similar pages](#)

[text-mining.org](#)
Provide a web home for people interested in **text mining** related technologies, with a mailing list and a resources section.
www.**text-mining.org**/ - 31k - [Cached](#) - [Similar pages](#)



Understanding the big picture





More applications in other domains

- **Automated grading of student essays**
- **Finding the best translation of a document into another language**
- **Finding synonyms (including in different languages)**
- **Estimating how well different parts of a policies and procedures manual are covered by a training course**
- **Finding documents where a certain sentiment is expressed**



Possible applications for iSTAR

- **evaluation of goals and objectives for each consumer**
 - how well do goals and objectives align?
 - does the text reveals anything about whether goals and objectives are being met?
- **identifying similar objectives (for same or different consumer)**
 - opportunities for ‘lessons learned’?
- **identifying high-level trends**
 - are there emerging issues?
 - could organizations position themselves to address these?
- **clustering consumers, according to e.g. shared or similar goals, diagnoses**
 - possibility for improved or more efficient assignment to case workers?
- **comparison of medical histories**
 - opportunities for ‘lessons learned’?
- **enhancing search capability within iSTAR text**
 - allow case workers to have relevant information more easily at their fingertips
 - enable them to perform their job better



DISCUSSION