

*digital*

Web Mining

Fool's Gold or Uncut Diamond?

In the early '90s, so just twenty years ago, Tim Berners-Lee was running the first web server at CERN in Switzerland. Since then the internet has grown into a data repository of enormous dimensions, and keeps on growing every second.

With the emergence of Web 2.0 and all manner of user-generated content – ranging from blogs via social networking websites to product ratings and discussion forums – the idea of harvesting all this information in some reliable and scalable form becomes very appealing. Consumers will talk about product experience on the web but they also use the web as an information source when planning a purchase. When it comes to trusted information about products, other consumers' recommendations **tend to clearly beat** the manufacturers' web sites.



So web mining - the exploration and analysis of relevant user-generated web content - is of great interest to marketers and market researchers.

A key output of web mining activities is the share of voice (SOV) a brand or product gains in relation to its competitors. SOV is usually based on counting how often a certain query appears across the observed parts of the social web in a specific time interval. The descriptive SOV is often enhanced by adding the element of positive, negative or neutral sentiment to the analysis. Even further significance can be added by assessing the influence of the user posting comments on a specific brand. Modern web mining systems also often allow researchers to engage with selected bloggers or users. This gives brand managers the opportunity

to move from passive monitoring towards active participation. Advanced analytical approaches go further than basic sentiment analysis and try to identify the themes of postings and conversations. The overall trend in web mining is moving towards 'understanding' what is being said.

The attempt to answer the ultimate question - whether web mining is just fool's gold or should be considered an uncut diamond - requires a closer look at the underlying technical procedures. The process of a web mining project is highly iterative and has five constituent steps: the **discovery** of relevant sources, the **extraction of data** from these sources, followed by **extensive data cleaning, analysis** and finally the **generation of insightful reports**.

The discovery stage is handled in various ways, but the two most popular approaches can be characterised as wide – covering as many sources as possible – or focussed. A key element of the discovery stage is the definition of appropriate queries. The



latter can go terribly wrong if the brands or products of interest have many semantic neighbours or twins. For example, the 'Mars' chocolate bar; Mars is a planet, a Greek god and a chocolate bar. Which one a consumer refers to can be very hard to identify, even if sophisticated Boolean logic is applied.

The challenge of query definition is to reduce the false hits, whilst at the same time capturing all the buzz about the product of interest. Assuming the discovery phase goes well and all data has made its way to the researcher's lab, it's still a challenge to remove irrelevant information. The blog posts and web conversations still have to be separated from all the noise that comes with them.

The analysis phase is then all about understanding the sentiment in which a brand is mentioned: positive, neutral or even negative. Getting this right is not an easy task either, since slang, abbreviations (like LOL), irony, sarcasm, complex sentences and other difficulties can lead to misclassifications. A

major challenge to analysis is the fact that the vast majority of mentions may all turn out to have the same sentiment – be it neutral, positive or negative. Factoring all these challenges in, there is still good reason for optimism. However, web mining as a valuable tool for marketers comes at a price. First of all, quality control is the key! It is very risky to draw conclusions from fully automated systems that come without various quality controls – preferably in the form of experts checking the results of each process step, refining queries, optimising the valence analysis, etc.

Depending on the objectives of a project, a decision has to be made as to whether broad “listening” to the social web or targeted “monitoring” of specific sources (e.g. recommendation sites) is the right approach to take. Last, but not least, whilst web mining can provide a very useful supplementary analysis, it should never be confused with research

data derived from representative samples. Although it may sound odd, given the size of the social web, for some categories for which there’s still just not enough data out there to provide robust insights. If expectations are realistic and objectives are clear, a thorough process makes web mining the uncut diamond but it can easily turn into fool’s gold, if not managed properly.

For more information please contact:



Norbert Wirth
Global Head of Innovation
GfK Custom Research
norbert.wirth@gfk.com

This article is from the series *'The Digital Connected Consumer'*. GfK Custom Research provides insights into traditional and new digital markets and lifestyles utilising state-of-the-art techniques. GfK Custom Research is a leading global fact-based marketing research consultancy covering more than 100 countries. www.gfkcr.com