Final Project Proposal: Finding and following memes

Martin Hecko

2008-02-12

Introduction

The concept of a "meme" has been invented by Richarch Dawkins and briefly discussed in his book "The Selfish Gene." If a gene is roughly a unit of biological inheritance then meme, as Dawkins writes, is a "unit of cultural transmission" (192). There has been some wrangling since Dawkins' brief mention of the concept as to what exactly meme is and if it is even a concept that can be scientifically studied. After all if gene is physically represented by DNA, what is the physical manifestation of a meme?

Description

While the topic of memetics – the science of memes – is a deep one I want to focus this project on a narrow aspect of memes. For this purpose I am going to assume that a physical manifestation of a meme is textual. This representation has some advantages: fixed form, ease of mass processing and availability of primary sources.

I would like to be able to identify which memes are popular and and how far have memes spread. In particular I would like to see how memes spread time and which channels they take most often.

Details

To track a meme's evolution it would be useful to have documents that are from the same source that changed over time. A Wikipedia complete edit history might be suitable for such analysis. At each point in time one could generate a list of n-grams (word sequences) that persist edits vs. word sequences that don't survive edits. Only limited amount of such sequences could be examined due to size and computational constraints – perhaps the n-grams could be selected based on their statistical probability compared to a standard English language frequency analysis.

The resulting data could then be visualized as an animation over time, where each phrase is represented by a pixel an a screen and changes brightness as the screen would show the frequency of the phrase during some time span. It would be interesting to identify phrases that have a long staying power vs. phrases that wink in and out of existence but never disappear.

Feasibility and Technical Considerations

Wikipedia dump is already available, but it doesn't include historical information. For this project to make sense a complete dump of Wikipedia would need to be obtained, although a smaller scale test could be conducted.

Since the algorithm would mainly involve text matching, no additional software other then the Hadoop cluster would be needed.

A suitable visualization environment such as Processing (http://processing.org/) could be used to construct the final animation – perhaps even with interactive elements..