Applying Random Surfer Model to Peer-to-Peer Network Distribution

Author: Dan Petrovic, http://dejanseo.com.au

Digital information preservation is a hot topic and a fertile ground for many bubbling solutions and models in both practice and theory. One of the emerging issues revolves around the fact that there is more information being produced today than we're able to store and analyse, not to mention attempts at prioritisation and archiving.

Kate Cumming (Future Proof - State Records, NSW) writes about this issue in her article titled: 'Here comes everybody': What does information ubiquity mean for the archives?

Two issues in particular caught my attention:

> *"The gap between what we create and what will actually have the capacity to keep is also growing exponentially. 'In fact, the production of digital information has already outstripped global server capacity by an estimated factor of four or five'"*
>
> *"We are also yet to fully define and understand the true costs associated with archival storage in this digital world."*

It seems clear that if we're to reduce administrative costs of information management we must minimise (manual) human intervention and secure continuity through robust and timeless protocols which are time, place, machine and platform agnostic. What about the problem of rapid information production? How can we process (capture, classify, verify, validate and order) the incredible amount of new information being generated every second?

Assessment of Existing Technologies

One fascinating phenomenon to observe is series of attempts by contemporary search engines (both specialised and general) to organise the information in a meaningful and useful way. We can safely say that none have solved the problem of real-time and social information stream appropriately, but one search engine has managed to order its results to a satisfactory level to say the least, and that is of course Google. The success of the world's biggest search engine lies in its ability not to only provide relevant results, but for its unique sorting criteria.

The original PageRank algorithm paper [1] was first published In 1998 by Larry Page and Sergey Brin of Stanford University with the promise of "Bringing Order to the Web". The basic principle of the algorithm is based on "The Intentional Surfer Model" [2], also known as "The Random Surfer Model". Origins of the basic logic for PageRank can be traced back to 1940s [3] and various scholars and scientists thought about the problem in various periods of the 20th century. The proposed algorithm observes a collection of documents and analyses how they link to each other, counting each link towards a page as a vote. Google uses this technology even today, though on a more sophisticated level [4] and in combination with other reputation signals. One can argue that documents which earn more trustworthy links and mentions tend to be more valuable than the ones nobody talks about. Although not bullet-proof, this method is capable of moderation of vast quantities of documents without any human intervention.

Could a similar voting system be applied in digital preservation frameworks to flag document in a digital collection as more valuable on the basis of its references?

Is Existing Technology Applicable?

After discussing the idea with Adrian Cunningham, the director of Digital Archives Program (Queensland State Archives) it was brought to my attention that the application of this algorithm may be more useful on an aggregation- rather than document-level.

> *"Archivists, for the most part, manage records in aggregations and not at the level of the individual document - that is partly because there are too many documents and too few archivists, but it is also because archival philosophy assesses the evidential value of records in their context and their context is largely derived from the aggregations of records of which they are a part and the inter-relationships between records within those aggregations."*

Another very good point is that archivists assess the evidential value of records from their context rather than their informational content. For example:

- Why was the record created?
- Who created it?
- During what activity?
- In which sequence of events or transactions?

Content analysis algorithm may not prove so useful unless they can extend to cover context and although some context can be discerned through content, this provides no reliable method on a broad level. Algorithmic treatment may work on a slice of the documents which tend to already have good recordkeeping meta data associated with them as that would provide good documentation of context. To make the situation even more difficult we can still add the challenge of standardisation of meta data to provide comprehensible and reusable format.

Creative Storage Solutions

Let us for a second imagine that by some magic the sorting and prioritising part no longer represents a challenge. What other problem do we face today? Storage of course.

Many talk about the cloud as an ideal solution, although in my view, it doesn't really bring anything revolutionary into distributed computing model - other than a catchy name. Many successful distributed storage platforms have evolved in the last decade and most of them were used for illegal purposes. If we are to learn one thing from movie and music piracy is that P2P (Peer-to-Peer) networks do provide robust and resilient storage solution. Many of these protocols (such as torrent) exhibit helpful features such as integrity checking, redundancy, fragmented storage, basic resource prioritisation, availability indexes, trackers and various other meta data.

At first the idea of algorithmic treatment of archived collections in P2P environment seems a bit farfetched, but there are some interesting developments in that area already. One of them is the LOCKSS program [5] which stands for (Lots of Copies Keep Stuff Safe). LOCKSS is an international community initiative with the objective of securing affordable preservation of digital content for libraries. They have been running for over a decade now and thanks to their open-source approach, developers are able to aid the evolution of supported platforms and applications.

The question remains, could a sophisticated algorithm with enough contextual data self-organise large-scale document archives in a distributed storage network without any human intervention?

References:

[1] http://infolab.stanford.edu/~backrub/google.html

[2] http://www.stanford.edu/~agupta03/Colloq.pdf

[3] http://www.technologyreview.com/blog/arxiv/24821/

[4] http://dejanseo.com.au/relationships-in-large-scale-graph-computing/

[5] http://www.lockss.org/lockss/How_It_Works