ISSN: 1138-9761 / www.cac.cat

QUADERNS
DEL CAC

Overcoming Babel: social mediation and intelligent systems in discovering, filtering, accrediting and personalising digital content

Enric Plaza

Head of the Department of Learning Systems at the Artificial Intelligence Research Institute (IIIA-CSIC)

enric@iiia.csic.es

Abstract

The convergence of digital content is transforming the distribution model from the centralised distribution of content to a more symmetrical model of network communication. This transformation also affects the production of content, this now being within the scope of any citizen with a computer and internet connection. The so-called Babel objection criticises this democratising effect. In this article we analyse the different mediation processes that relate content with recipients that are present both in the centralised distribution model as well as in that of network communication. The aim of this article is to show that it is viable to develop the discovery, filtering, accreditation and personalisation processes of a network communication model where consumers are also contributors.

Key words

Artificial intelligence, personalisation, search, mediation.

Resum

La convergència de continguts digitals transforma el model distribució: d'un model de difusió centralitzat de continguts a un model de comunicació reticular, més simètric. Aquesta transformació també afecta l'elaboració de continguts, que és a l'abast de qualsevol ciutadà amb un ordinador i connexió a internet. L'anomenada objecció Babel *critica aquest efecte democratitzador. En aquest article analitzem el diferents processos de mediació que relacionen els continguts amb els destinataris i que són presents tant en el model de difusió centralitzat com en el de comunicació reticular. L'objectiu de l'article és mostrar que és viable desenvolupar processos de descobriment, filtratge, acreditació i personalització en un model de comunicació reticular on els consumidors són també contribuïdors.*

Paraules clau

Intel·ligència artificial, personalització, cerca, mediació.

Introduction

The convergence of digital content is transforming the distribution model from centralised distribution of content (from few centres to many users) to a model of network communication (from many to many). This transformation also affects the production of content, this being within the scope of any citizen with a computer and internet connection. In principle, the network communication model is symmetrical, in the sense that any node can be both consumer and creator of content at the same time, be it data, information, knowledge or culture. This democratising effect has been criticised by the so-called *Babel objection*: if everyone can talk, no-one can listen because of the resulting cacophony (information overload). If the Babel objection is right, democratisation will fail and internet citizens will stop being active contributors and become passive consumers. However, if a schema can be organised that efficiently and easily relates content and its recipients, we will be able to overcome the Babel objection.

In this article we analyse the different mediation processes that relate content with its recipients, i.e. discovering, filtering, accrediting and personalising. These processes are present both in the centralised distribution model as well as in that of network communication, the latter merely adding a quantitative difficulty in carrying out these processes. The aim of this article is to show that it is viable to carry out processes of discovery, filtering, accreditation and personalisation in a network communication model where the consumers are also contributors. In particular, we will analyse two basic elements: *a*) information content provided by contributors themselves on mediation processes, and *b*) the use of artificial intelligence techniques in handling large amounts of data in discovery, filtering, accreditation and personalisation processes.

Network symmetry and ownership of the material means of production and distribution

The transfer involved in any change in paradigm - currently the transformation from a distribution model (from few to many) to a network communication model (from many to many) – leads to two kinds of opposing responses: the response given from

an apocalyptic perspective and that from an integrated perspective. Umberto Eco (1964) characterised these two opposing theses (apocalyptic and integrated) with regard to the mass media of the 1960s and today we can detect some similar responses. On the one hand, that of the apocalyptic/reactionary perspective, that sees only problems in the new paradigm of internet information: cacophony, information overload, lack of credibility, etc. On the other hand, that of the integrated/revolutionary perspective, that stresses only the positive possibilities: better access to information, democratisation of the information distribution process, more potential for criticism/monitoring actions of the groups established, ease of coordinating large numbers of people, etc.

The answer is not the happy medium but accepting that there are both negative and positive aspects and analysing how we can help achieve these positive possibilities and with what mechanisms, and how we can do away with the negative effects. Technology is not neutral in this respect, nor is the legislation that limits its possible options: the mechanisms employed may destroy some of the positive possibilities or preserve some of the more negative effects.

For this reason we must first analyse the effects of the technological change not only in the sense of society and customs but also in economic and productive terms. From the most abstract point of view, this change in paradigm gives rise to a medium more similar to the telephone network (where everyone can communicate with everyone) than to the model based on publishing firms/content providers. Symmetry is a characteristic of the network structure: all the nodes are equal members of the network, all receive and transmit content. This symmetry can also be found on the network of networks, the internet, but this is not enough to explain the change in paradigm. The second factor is the personal computer that, unlike the telephone, is a medium for creating, elaborating and producing content (be it data, information, knowledge or culture) and is particularly a highly decentralised medium of production, i.e. owned by individual citizens and not by companies or the state.

It's the combination of the digital production medium (computers) and the digital distribution infrastructure (internet) within a context of decentralised ownership that transforms the political economy from an industrial information economy into a *networked information economy* (Benkler 2006). A historical example of economic change is the cost of creating newspapers at the start of the industrial economy era. According to Benkler (2006), starting up a new newspaper in the United States in 1835-1850 initially cost 10,000 dollars (in today's terms), a cost that went up to 2.5 million dollars (in today's terms). This sharp change in costs wiped out an ecosystem of small newspapers with different kinds of organisation and funding (with a weekly circulation higher than Europe in a United States of only 17 million inhabitants).

According to our experience, gained under an industrial information economy, it seems that the only two alternatives for content production are (large) market-based firms and state companies: it's difficult for us to imagine "serious" alternatives beyond these two models. In spite of this, the ownership and financial costs of producing and distributing content have fallen extraordinarily (computers plus internet connection). This is what Yochai Benkler (2006) calls "social production", which is added to market- or state-based organisations. Consequently, the ecosystem of creating, elaborating and producing content we can expect in the near future will be much more decentralised in comparison with the industrial system.

Discovering and filtering

Finding new content has always been carried out "formally" with guides and catalogues but also "informally" by using social networks: a friend or acquaintance tells us that such and such a radio programme plays music we might like. The internet has added the proactive possibility for a person to use search engines (like Google) to find new content. It must be noted that the first proposal for discovering content was "formal" and developed by Yahoo, attempting to make a website guide/catalogue. This catalogue was carried out manually and was not scalable because of the large number of websites in existence.

The alternative was to use web search engines, applications based on information recovery techniques adapted in order to analyse, index and recovery websites, e.g. Aliweb in 1993 and Altavista in 1995. Today Google is the most popular search engine but we must analyse the technological reason for its success: the analysis and use of user-provided content (UPC). The central idea to the PageRank algorism used by Google is based on an analysis of particular content provided by the user: hyperlinks that relate two websites. In effect, the user declares that (the content of) the page he or she is writing is related to (the content of) the pages it is linked to. PageRank analyses the network of relations provided by users as links to assign to each page P a specific degree of importance determined by (the importance of) the pages $P_1 \ldots P_n$ referring to page P. This algorism is based on previous work carried out in bibliometrics on citation analysis: the innovation of PageRank is that it focuses on the analysis and exploitation of a specific kind of UPC, hyperlinks, to filter or distinguish more "important" content from less "important" content.

The techniques of artificial intelligence can improve discovering and filtering processes within the context of the so-called Semantic Web. The Semantic Web, proposed by Tim Berners-Lee, the creator of the first website, is based on the "annotation" of web content using ontological terms, so that the content produced by humans can be understood by automatic intelligent systems. However, this new web technology is "sectorial": each sector requires its own ontology (a formal description of the meaning of the terms used in this sector). For example, content of a legal nature would have a legal ontology defining terms such as *fraud*, while content of a medical nature

would need a medical ontology. With regard to musical multimedia content (<http://musicontology.com>) this is the most developed at present and the BBC has started to apply it to its website.

Another way of improving discovery and filtering is to analyse the behaviour of user communities when they search and to learn to filter more intelligently, so that we can discover which content is really interesting for that community. University College Dublin is working on this area: instead of developing an ontology for each theme, the system learns by observing what user groups interested in football or photography or iPods do. The techniques employed are similar to those of recommendation systems, like the simple but well-known systems to recommend books on Amazon or music on AppleStore. Analysing the actions of users, when discovering and selecting what they are interested in, provides a much more personalised result for each user.

Accrediting and personalising

While discovering and filtering are mainly concerned with the relevance of certain content for the user, a second dimension that is also important is the credibility of the content and the reputation of its origin (or sources). Without doubt, the supposed "lack of accreditation" of content, in addition to the large amount of information, is one of the most important factors within the pessimistic opinion concerning the Babel hypothesis. This pessimism concerning the possibility of a decentralised, efficient mechanism to distribute content comes from the model established by the large mass media, where these big organisations consider it their role to classify content into hierarchies, for example which content is for the front page and which should have a small or zero space allotted to it. In this model, the large number of organisations provides both diversity of hierarchy and accreditation of content (based on the reputation of the organisations). However, a criticism of the current situation is clear: the number of mass media organisations is small in order to guarantee diversity, and content is often published without much comparison with reality for reasons of immediacy.

From a citizen's and user's point of view, the accreditation provided by the mass media is quite relative: there are people who trust certain organisations and not others. This trust is due to the reputation models assigned to specific organisations and people. To overcome Babel, it is therefore necessary to create and maintain systems that can evaluate the reputation of content authors/distributors via decentralised mechanisms that replace the hierarchical mechanisms of the mass media.

Given that social reputation and accreditation are also information goods, both can be treated like any other content. Social reputation and accreditation can therefore be created in a decentralised way by the very users/producers/consumers themselves (UPC). In fact, one example of this is the website Slashdot (<http://slashdot.org>), which allows precisely this and has become, for the moment, one of the technological news bulletins (*News for Nerds*). Its operational principle is very simple: users provide the URL with a news item or content in general and add a comment regarding its interest. Other users also add comments, which often run into the hundreds. Slashdot uses *ex post* peer reviews to evaluate the credibility or quality of the comments. This method is a variation on the system of scientific publication (peer revision prior to publication), in which the revision is carried out *a posteriori*.

Slashdot does not try to stop irrational or erroneous content from being published but merely compares it with elements that corroborate or refute it. Habitual users accumulate "karma points" for their good actions (or have points docked for bad actions). Consequently, a reputation mechanism is created, neutrally and automatically, that helps users to weigh up the alternatives in conflictive situations. The result is the ordering of content, i.e. a hierarchy, which has been produced, however, in a decentralised way by the very community of those interested in technological news and content. Research is currently being carried out into more sophisticated reputation models at our Artificial Intelligence Research Institute (IIIA), among others, with the aim of creating far-reaching accreditation platforms.

Finally, personalisation is typically a process that relates certain content with the affinity (interests or preferences) of a user. One of the most widely used techniques is collaborative filtering, used for example by Amazon to recommend books, films and, as also done by AppleStore, music. Collaborative filtering makes a prediction regarding the elements that might be most closely related to a person, comparing the elements that are related to other "similar" people. The way to determine that two people are similar may vary, but essentially the registered behaviour of users is compared (in the case of Amazon or AppleStore, the elements bought by each person). Apart from this technique, there is currently quite a lot of research to develop more closely adjusted recommendation systems. For example, a spin-off company of the IIIA, MyStrands (<http://www.MyStrands.com>) develops social recommendation technologies particularly in the world of music. Recommendation and personalisation systems are a new and very active field within artificial intelligence, with the first congress held in 2007, and they are likely to become established in the near future as a technology as ubiquitous as content searches today.

Conclusions

The processes of decentralisation and automation that act on the discovery, filtering, accrediting and personalisation of content will certainly have consequences we cannot predict, but to end I would like to mention the importance of the phenomenon known as "the long tail". This term was coined by Chris Anderson (2006) to argue that, in the new internet cost struc-

ture, products with few clients or sales, jointly, could achieve a greater market volume than products with more clients or sales. These curves are known in statistics as Pareto tails but are often called 80/20 curves in mail order sales. This means that 20% of the products account for 80% of the sales and "the tail" is the remaining 80% of the products, which account for 20% of the sales. Current studies show that, on the internet, this curve becomes 72/28, a considerable change in practical terms. So, for example, Amazon can have an extensive catalogue that includes a lot of products with relatively low sales, i.e. niche products, but which, as a whole, generate a large part of its business.

This is relevant because the so-called "fragmentation" of content is a phenomenon that will continue to grow due to the long tail effect: increasingly more content will be created for niches, i.e. for markets that are not mass markets. The mass media is currently changing into a myriad of services and content aimed at medium or small-sized interest groups and this will continue due to the action of new technologies and cost structures. Those with an apocalyptic view may fear Babel but I have attempted to show that there are ideas and techniques that can organise this new internet galaxy in a new, decentralised and social way. However, uses and habits will change and, admittedly, this will lead to anxiety. I personally believe that nostalgia for the time when we used to all watch the same film on a single TV is mistaken.

Note

1 For an example of the use of ontology in searches see <http://www.cognition.com>.

Bibliography

Benkler, Y. *The Wealth of Networks: How Social Production Transforms Markets and Freedom.* Yale: Yale University Press, 2006.

Eco, U. *Apocalittici e Integrati*. Milan: Bompiani, 1964.

Anderson, C. *The Long Tail: Why the Future of Business is Selling Less of More*, New York: Hyperion, 2006.