

Research Report

March 2008

Becta leading
next generation
learning

Emerging technologies for learning

Volume 3 (2008)



'If it quacks like a duck...'

Developments in search technologies

Emma Tonkin UKOLN

Digital content is widely viewed as a viable replacement for traditional educational resources. Projects such as One Laptop per Child place interaction with digital content directly in the centre of a constructivist approach to education. There is a widespread sentiment that 'Google generation' children are sufficiently curious to educate themselves, given the ability to access information. How true is this? When we look at any new technology, we are usually too optimistic. Much as the internet itself in its early days was seen as a social enabler and a means to democratise society, the means to search and retrieve information is sometimes seen as a means to make knowledge available and therefore bypassing the intermediate steps of education.

In practice, however, there are complications, which can be identified by examining the capabilities and limitations of the technologies in question. There is a growing awareness that the high rank of a search engine result does not necessarily reflect the quality of the resource, and increasing concern that many are too quick to trust the ever-increasing information retrieved from the internet.



Traditional search engines are not all-powerful. For example, Google assembles evidence via links (the famous PageRank algorithm) and page content to index pages effectively. As such it is limited to pages that are *accessible* to its web spider, but many are not, giving rise to the *grey web* – publicly available pages that cannot be indexed. For the search engine words are simply abstract symbols with no underlying network of concepts, giving rise to a current interest in *conceptual*, *contextual* and *semantic* search, where pages are mapped against a structured set of terms or concepts. Multimedia content presents an additional set of practical difficulties, for the pragmatic reason that video streams are more difficult to de-construct into unitary features than a page full of text.

Search engines are not one-size-fits-all. Users vary greatly in needs and abilities through age, background, interests and task. Hardware form factors, abilities and interface peculiarities are significant factors in designing appropriate interface query modes. Sometimes a search is not what is needed at all, but discussion, browsing, or informal interaction with a peer group. A vast array of methods and technologies exist that, separately or in combination, allow search engines to respond to users' individual needs and circumstances. In this article we explore some factors involved in personalisation of search results, and several means of organising and searching through information: the Semantic Web; social tagging; web, multimedia and mobile search.

Digital objects without semantic annotation are just ones and zeros on a hard disk, just as a library without a catalogue is merely a stack of books. Some information is needed to enable the computer to locate objects usefully, but how much detail is required? Is this information extracted from the object or contributed by a human? Technical concerns abound; how should the data be encoded, and an object *annotated* with an accessible representation – *metadata*, data about data, relevant facts such as author, title and publishing date, or *relations* to other objects? Various mechanisms exist that provide a structured means of semantic annotation, including metadata standards such as Dublin Core (DC) – and the Semantic Web (SW).

The Semantic Web

Laying aside technical details, a computer needs some information about an object before it can recognise it as the resource that we are looking for, from the very obvious ('it's an image') to the detailed ('it's a picture of Alexander Graham Bell, who invented the telephone'). There is a reference problem associated to describing a digital object in terms of a number of entities and objects, 'telephone' and 'Alexander Graham Bell'; what are these terms? What do they mean? How are they related?

The set of concepts and technologies collectively referred to as the *semantic web* originated in 2001 with a landmark article written by Tim Berners-Lee, James Hendler and Ora Lassila (Berners-Lee *et al.*, 2001). Instead of providing unstructured records and then describing them separately, or extracting semantics from the object itself, the semantic web holds its own metadata. Data is published in a machine-readable form, so that a computer can interpret and apply the information available on the Web. The computer can then perform sophisticated tasks for the user.

For example, the computer, which lacks the commonplace knowledge required to reason about Alexander Graham Bell and the telephone, learns to relate concepts using well-formed facts, entities and relationships in order to reason about the world:

Alexander_Graham_Bell (is_a) man

A man (type_of) person

Alexander_Graham_Bell (inventor_of) telephone

telephone (type_of) communications_technology

The computer responds to queries about the invention of communications technologies by offering a resource depicting Alexander Graham Bell. Perhaps if it has enough records, it will also mention Elisha Gray, whose patent application arrived just after Bell's.

Berners-Lee's vision included more practical tasks. As such, the example scenario described two siblings, Pete and Lucy, arranging physical therapy sessions and chauffeuring duties for their mother's prescribed treatment. Instead of doing the research and scheduling the tasks themselves, Pete and Lucy each task their software agents (software that acts in the interests of a user, acting as their agent in a transaction) with researching the various *providers* who are able to offer the *prescribed treatment*, which are *in-plan* (within the expense budget), within a *20-mile radius* of their mother's home, and which are rated as *excellent* or *very good*. The agents then negotiate *appointment times*. Terms italicised in this paragraph are recognised by the agent as semantically meaningful.

Those who have chosen a doctor or healthcare provider for themselves or a dependent may not recognise themselves in Pete and Lucy. What are we really willing to automate based on a ranking score? One answer is that the process is not fully automated. The device retrieves recommendations based on a semantically marked-up subset of the data available on the internet about the various topics of interest. Moreover, if Pete or Lucy wish to check any review text that might accompany the various recommendations, they are welcome to

do so – but *sentiment analysis* of a text is a notoriously difficult natural-language programming problem, so they would be unwise to depend entirely on their agents' judgement in that area.

To a cynic, this scenario demonstrates nothing more than can be done using a set of databases. Travel agents have been able to recommend holidays according to distributed database records for many years now. This is true, but the point here is that arranging public access to those databases is not required. The data from these databases is published to the Web, where Lucy's agent can find it and use it as a basis for reasoning.

The Semantic Web in practice

The discussion surrounding the Semantic Web (SW) is one of extremes; a somewhat illusive, if keenly felt, dichotomy has been established between those who favour structure and those who feel, as Wittgenstein did in his later years, that concepts defy wholly objective description. Discourse is heated and often bitter, a warning sign that something far older and closer to the heart than technical concerns motivates the debate. The SW is frequently heralded as the dawning of a new era and has received substantial funding. Yet although a great deal of research is published every year, our talkative cynic might remark upon limited practical implementation outside the research community.

Like many ambitious projects in science, the Semantic Web has many detractors; as with superstring theory, there are few results that showcase the utility of the approach relative to alternatives. One high-profile critic, Clay Shirky, caused a flood of indignant responses with his 2003 article (Shirky, 2003), that described the SW as 'a machine for creating syllogisms'. This was inflammatory. Indignant commentators pointed out technical flaws, such as the point that the SW is largely based on first-order predicate logic (FOL), not syllogisms. Areas of life in which such logic is widely and successfully applied were held up to view. Yet Shirky's criticisms had many valid components. For example, his frustration with the level of complexity of the standards has been widely echoed by developers, and it is certainly true that there exist aspects of human behaviour and understanding that are not easily modelled in logic programming.

A quantitative analysis of the amount of Semantic Web pages on the Web from 2005 found that the majority of data available in the flagship standard of the effort, RDF (Resource Description Framework) is simply data encoded in RSS, the newsfeed format variously defined as Rich Site Summary and Really Simple Syndication, amongst others (60 per cent – versions 1.0 and 0.90 of the RSS standard are based on drafts of the RDF standard). A small amount is FOAF, or Friend Of A Friend, another technology based around RDF that was poised to take over the world in '03; social network applications such as LiveJournal, Facebook, Typepad and MySpace are the inheritors of the crown. Some make



use, incidentally or otherwise, of FOAF, but it seems that centralised storage of data has sufficed for social networking applications up to the present date. Only a minority of RDF records online represented novel SW developments.

Creation of a reasonably accurate model of actual experience – as opposed to an idealised model – is a difficult task. To quote McCool (2005), the ontological data model makes representation of any nontrivial factual information difficult because it can't represent context. Artefacts such as social networks (Paolillo *et al.*, 2005) are rather imprecise compared to the precise, static conception of meaning encoded into an ontology. Like any database, aging structured data needs maintenance. It is difficult to reengineer a system. Instead, we may resign ourselves to using an increasingly clunky and unintuitive system on the principle that it is a) familiar and b) still works.

As a rule-of-thumb, ontologies work well when they represent commonly held theories or models that are explicitly relevant to a given topic. For example, we apply a Linnaean taxonomy to categorising species, prioritising uniformity of classification over personal experience. One might find a platypus funny and a funnel-web spider horrifying, but subjective reactions do not form part of our formal classification of species and hence are understandably excluded from scientific discourse on the topic of Australian fauna. However, such nuances can be collected and added to our reasoning using other technologies, such as social tagging.

The role of AI and data mining

Who will generate SW annotations? There may be a tipping point. If organisation A makes its information available, perhaps organisation B will do the same, and eventually a critical mass of data will be met. But Berners-Lee expects the SW to be populated in part by means of automated approaches to information extraction from digital objects.

By 'semantic', Berners-Lee means nothing more than 'machine processable'. The choice of nomenclature is a primary cause of confusion on both sides of the debate. It is unfortunate that the effort was not named 'the machine-processable web' instead. This, along with some optimistic usage scenarios, is a primary cause of extraordinarily high expectations in those who take the term at face value, and hence to fierce criticism of what is seen as an overly-ambitious area of research. In summary, the SW is 'a webby way to link data'.¹ Those evaluating the technology are well advised to look at what it really represents today, rather than what it may one day become.



¹ <http://journal.dajobe.org/journal/posts/2007/03/17/semantic-web-is-webby-data>

Initially, the SW was often presented as all or nothing, revolutionary technology. Who can blame commentators for seeing the Emperor's new clothes in such an ambitious technology? At present, practical deployment remains fairly 'heavy' and complex, involving an investment that, though rewarded with a more explicit and powerful representation, is arguably not always necessary. Much of the software remains somewhat experimental, and the IT workforce has limited familiarity with specific concepts or technologies.

Many see a pragmatic need for a compromise, a *slightly semantic* Web. Later, we will examine a few technologies, recently proposed or resurgent, that form part of today's functional compromise.

The lower-case semantic web

Microformats

Accepting that, for some purposes, the Semantic Web may be too much of a good thing, various forms of lightweight semantic tagging have been suggested. For example, the microformat was developed as a simple method of making Web pages that are a 'little bit semantic'. Appropriate attributes are placed in XHTML in order to render everyday information machine-readable. Current availability of marked-up content and services suggests an encouraging future (Allsopp, 2006), though it is too early to say whether the approach will be widely adopted in the long term.

Figure 1: Hcard allows us to easily mark up human-readable text; for example, UKOLN's contact details:

<pre><h2>Contacting UKOLN</h2> tel: ++44 (0) 1225 386580
 fax: ++44 (0) 1225 386838
 email: ukoln@ukoln.ac.uk web: www.ukoln.ac.uk </p></pre>	<pre><div class="vcard"> <h2>Contacting UKOLN </h2> tel: ++44 (0) 1225 386580
 fax: ++44 (0) 1225 386838
 email: ukoln@ukoln.ac.uk web: www.ukoln.ac.uk </p></pre>
--	--

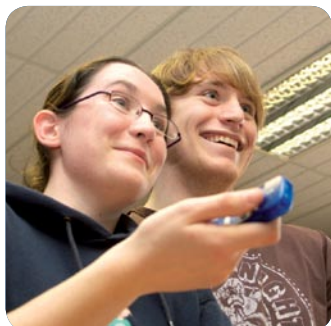
A microformat reader recognises this as a visiting-card, which could be stored in an address book. This minor change for the author of the web page allows the browser to recognise the data and how it may be treated, stored and queried. Microformats exist for a few popular formats (addresses, event descriptions and news items, for example), but much of the text that is placed on the Web ‘resists simplistic representations’ (McCool, 2005).

Intrinsic and extrinsic meaning

Another popular technology today is social tagging, which is much less about establishing what a data object intrinsically is, and much more about finding out what people think about it and what it means to them. Social tagging systems (Guy and Tonkin, 2005) allow users to apply short plain-text descriptions to a data object. Tagging systems contain elements designed for personal information management (Kipp, 2006), topic-based keywords that would not be out of place in a controlled vocabulary, and informal descriptions. Tags may be applied by an individual for their own use or for a community or global audience (Tonkin *et al.*, 2008) and can be thought of as digital annotations in the sense described by Marshall (1997). Several widely publicised and popular sites make extensive use of tags, including Flickr [<http://www.flickr.com>], in which users tag their photographs, and del.icio.us, a social bookmarking service in which any user may tag and share any resource available on the Web. Tag systems may also contain other types of information, such as geotags (referent location information).

LibraryThing [<http://www.librarything.com>] is a great example. Despite its use of constrained subject vocabulary, LibraryThing, like most uses of social tagging, is all about aggregating points of view. On LibraryThing, Library of Congress Subject Headings (LCSH) are applied for book categorisation, supplemented by additional LCSH categories provided by libraries across the United States – and users’ own opinions. The site allows users to click through and see book recommendations and user preferences. The downside of this connectivity is the ease with which young readers may find themselves outside their comfort zone, an effect common to any highly connected network.

The difference between intrinsic and perceived meaning in recent library history can be illustrated by Ray Bradbury’s statement that his famous novel, *Fahrenheit 451*, is ‘not about censorship’ (Johnston, 2007). LibraryThing [<http://www.librarything.com/work/4248>] suggests that the consensus of opinion does not support Bradbury’s opinion of his work – the tag cloud for the book (Figure 2) shows the terms most commonly applied to the resource. Tag clouds are a common visualisation method for tags, and give a quick visual idea of the sort of terms applied to an object.

Fig 2: The LibraryThing tag cloud for *Fahrenheit 451*

(42) 20th century(58) american(100) american literature(72) book burning(75) **Books(111)** books about books(16) Bradbury(46)
 censorship(234) classic(342) **Classics(171)** dystopia(512) fantasy(44) favorite(24) favorites(17)
Fiction(1,091) fire(15) future(61) futuristic(29) high school(19) literature(141) novel(155) own(61) owned(25) paperback(45)
 political(24) Political fiction(18) politics(31) Ray Bradbury(51) read(226) reading(15) satire(25) School(25) **sci-fi(348)** science
 fiction(971) sf(143) sff(15) Social commentary(43) society(17) Speculative Fiction(30) sbr(16) totalitarianism(29) unread(48)

Tagging is widely criticised as ‘noisy’, and the wide range of annotations – in particular intended audience and level of formality – attracts criticism from commentators. Tags cannot be trusted; they are explicitly points of view. On the other hand, tagging systems often form part of a social network (Tonkin *et al.*, 2008) that suggests each of us will find (or the system will recommend) taggers whose judgement we feel we can trust.

Spamming and gaming search

Can a model succeed in which semantic annotations are automatically considered trustworthy? These differing methods pit authority against managed chaos. Given present text-based search technology, we do not face a lack of information, though we may lack trusted sources. Search engines can be fooled. A lucrative industry exists around the process of malicious gaming of search results. We face a world in which almost anything, anywhere, can be found, a vision which Morville described as ‘ambient findability’. Most things are ‘findable’. In the near future, the problem of search may become subordinate to the problem of filtering information, applying technologies initially designed, in the language of ubiquitous computing, to *augment* reality, to enable voluntary perceptual blindness – allowing us to direct our attention effectively to the information that matters to us.

How do we search?

Software agents are most capable when data placed before them is formally encoded in terms of semantics that they can relate to their own set. They do not, of course, ‘understand’, but depend on formal reasoning and the set of symbols available to them.

An agent might search as follows:

‘Find: [Any document] [about] [unique topic ID]’

receiving the response:

[Document at <http://www.uniquetopic.com>] [about][unique topic ID]...

The agent has received a readable response, is satisfied and processes these documents.

Is this an accurate model of user behaviour? Grosky (2002) notes that management systems for multimedia document collections and their users are typically at cross-purpose. Systems normally retrieve multimedia documents based on low-level features; users apply the abstract notion of *relevance*. Though classical, this is by no means the ideal model of user behaviour during interaction with a search engine.

Instead, successful search methods and interfaces are often designed to support specific aspects of information seeking on the Web. We seek known pages or other objects frequently (Choo, 2000), but Rieh (2004) found that for most people, we turn to the search engine when we cannot think of an appropriate topic-specific site. Queries *evolve* as the search progresses, terms are refined, and perhaps various different search engines with different capabilities are tried – for example, Rieh found that the ask.com natural language querying abilities are commonly used for some types of query. Ask.com is able to answer questions such as ‘How cold is Pluto?’ using a feature called ‘Smart Answers’ that provides answers at the top of the page. A version called ‘Ask for Kids’ [<http://www.askforkids.com>] provides a user-friendly question-refinement system enabling young people to refine their search (‘Tell me about the President’. Which one? ‘George Washington’). Wikipedia and its Simple English counterpart [<http://simple.wikipedia.org>] provide an analogous service.

Bates describes this model of search and retrieval as ‘berrypicking’; the user tries a succession of search terms, retaining useful or interesting results. By contrast, in ‘serendipitous browsing’, information is often found by happy accident, through semi-directed search or wider reading around a topic. Precision and serendipity do not go hand in hand. The most authoritative piece of information about a topic may not be the best resource. What about an opinion piece, a rant or a reaction, a thoughtful review or a beginner’s introduction offering a new perspective? In multi-disciplinary academia, establishing links between groups is difficult to do without going back to the basics, questioning and re-evaluating what we think we know.

The optimal user, like the optimal student or researcher, is information literate, patient and able, ready to engage in an interactive process of negotiation – and he or she is rare. A recent CIBER briefing paper (2008) suggests that today’s digital library users have very heterogeneous approaches to information seeking. They ‘skim’, viewing a few pages from an academic site, then leaving. They spend little time on websites, but ‘power browse’, looking for ‘quick wins’, a finding supported by studies such as Nielsen (2006), which showed that users browse pages by scanning them unevenly, focusing mainly on the top left of the page. The CIBER study suggests that the ‘Google generation’, who look to the internet as a primary resource for information, have a low appreciation of what the internet is, little ability to determine the trustworthiness or relevance of resources, low information literacy and poor understanding of their information needs. The extent to which these findings can be generalised is unclear, but the

paper adds that there is ‘little evidence...that Google generation youngsters are fundamentally “different” to what went before’.

Lecturer Tara Brabazon (author of *The University of Google*) has banned her students from using Wikipedia and Google in their first year of study, stating that these services offer quick answers but that dynamic and critical thinking skills are required before they are suitable for students’ use. First-year students are instead provided extracts from peer-reviewed printed texts.

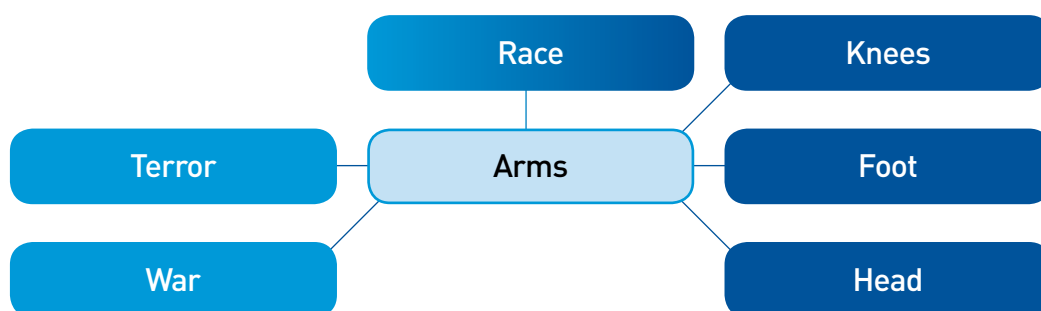
Data mining and artificial intelligence

The present state-of-the-art in artificial intelligence is not very good at understanding what things mean. However, computers are very good at brute-force calculation and abstract symbol manipulation. Anything possible in reasonable amounts of time by applying mathematical algorithms to a data structure is achievable by a computer.

Most of the statistical methods currently in use are based on some interpretation of the *distributional hypothesis*: that is, that two objects (such as words) are similar to the extent that they share contexts (Harris, 1968). That context might be adjacency within a sentence, co-occurrence within a sentence, paragraph, or document, or some other measure such as co-occurrence with another term.

For example, the words ‘for’ and ‘example’ frequently occur side by side. A document which frequently mentions ‘search’ might also contain frequent uses of ‘engine’, or possibly ‘rescue’, depending on the actual theme of the document.

Automated approaches can also distinguish between multiple uses of the same term. This is a major challenge for search engines such as Google: how does one distinguish between the use of the term ‘arms’, as in ‘arms race’, and its use in the sense of ‘arms, knees and head’? (Windows, 2004). This is actually one of the easier problems to solve via the distributional hypothesis; the use of the terms looks rather like this:



Where the word ‘arms’ co-occurs with ‘terror’, warfare may be involved. Where it co-occurs with ‘foot’, it probably refers to appendages. The role of statistics in guessing at meaning is in building useful generalisations based upon the available evidence. Given initial choices from both clusters, the reaction of the user may

permit the search engine to tailor further results according to the definition that has proven to be of interest. If they are exclusively interested in the topic related to warfare, results could be biased towards that definition. A similar approach can enable *entity disambiguation*, the ability to guess whether a particular use of the term 'George Bush' refers to the father or son. Co-occurrence is valid for words, but also for information like Web addresses and references to images or video.

There are many different approaches to statistical analysis. The mature technologies in use are not closely linked to human intelligence or judgement, though they may be inspired by research in those areas. The spam filter is probably the best-known example of supervised learning, based on a statistical trick known as Bayesian classification, described in Figure 3. Bayes does not return certainties, but a measurement of probability that a given object can be classified under a given label, given previous experience in the area.

Commercial search engines do not generally learn in the manner of a Bayes classifier, but work instead from ranking methods beyond the scope of this article. However, they too answer according to calculated probabilities: 'this query is similar to that article'.²

Figure 3: Bayesian reasoning in classification

Bayesian reasoning is probably the most familiar form of automated classification. It is mainly seen in spam filters. It is based around the following observation (Bayes' Rule):

$$P(\text{Cause}|\text{Effect}) = \frac{P(\text{Effect}|\text{Cause})P(\text{Cause})}{P(\text{Effect})}$$

This cryptic equation can be read in English as: (the probability that something that quacks is a duck) equals (the probability that ducks quack multiplied by the likelihood of there being a duck) divided by (the likelihood of hearing something quack). If it quacks like a duck and we have reason to expect to see a duck – and we have ample experience of the ways in which ducks quack to draw upon when making our judgement – then we correctly classify it as being a member of the genus *Anas platyrhynchos*. The advantage of Bayes is that it provides us with a quick and easy way of allowing a computer to make similar judgements. Search engines use many approaches, but in general the philosophical aim is similar: to determine, as cheaply as possible in terms of memory and computational power, the characteristics of the various objects on the web, the links between them, and their relationship to natural-language queries.

² Two excellent books for further reading in the mechanics of search engines are *Understanding Search Engines* by Michael W. Berry and Murray Browne (2004), and *Geometry and Meaning* by Dominic Widdows (2004).

Multimedia search

The distributional hypothesis of semantics is not limited to text. For example, when an individual browses the Web, he or she travels on a *browsing path* from one page to another. Grosky (2003) suggests that two adjacent pages on that path of clicked links, particularly the part of the page surrounding the link that takes you to the following page, are probably similar in topic. Pages with similar layout are often assumed to be similar in topic; text elements that have similar layout may have a similar function, etc.

With data mining, we extract relevant features from a digital object, such as the key signature or melody of a piece of music (Hartmann *et al.*, 2007), the prevailing opinions in customer reviews (Hu *et al.*, 2004), or warning signs in lung cancer diagnosis (Perner, 2002). We can look for patterns and similarities between pages, or cluster web pages or other documents according to feature, such as the appearance of a 'sky-blue' colour in images.

Specialist search engine technology is moving ahead. In audio search, Nayio [<http://www.nayio.com>], though as yet unimpressive, holds out the promise of eventually being able to identify tunes by humming the theme into your microphone. Various video search services provide methods of searching by keywords spoken throughout the film. Google Video (Beta) searches the closed captioning of various television channels. Speechbot, a project run by Hewlett-Packard between 1999 and 2005, indexed over 17,000 hours of radio programming via speech recognition, but the project was terminated with the closure of HP's Cambridge research lab. Blinkx (founded in 2003) searches video files using a similar method, offering services to MSN, Live.com, Lycos, Infospace and AOL. PBS runs several similar archives, also using speech-recognition technology, that has been applied to index several PBS programmes over the last several years, such as News Hour [<http://www.pbs.org/newshour/video>] and the Mathline series of mathematics educational videos [<http://www.pbs.org/teachers/mathline/lessonplans/aboutvid.shtml>].

TASI's review of image search engines [<http://www.tasi.ac.uk/resources/searchengines.html>] primarily describes image search methods that rely on the text surrounding the image for information about the image content. Google's web spider (googlebot) does not analyse images directly. By comparison, picsearch.com searches on information taken directly from the image such as dimensions, file size, file type and colour information, all features of the image itself rather than the image context within a document. This approach has also been taken in video indexing, such as for example by Clippingdale and Fujii (2003) and Lee (2005); hybrid or *multimodal* approaches that index by combination of video and audio information (for example: faces and voices) have also been proposed in the research area (see for example Albiol *et al.*, 2004). IBM's 2006 Marvel search service (for local installation) uses advanced content analysis techniques for labelling and analysing data, and

provides hybrid and automated metadata tagging. Recent UCSD research, a supervised content-based image analysis for recognition of image elements (Carneiro *et al.*, 2007), demonstrates the complexity of the problem. Recognition of simple objects is shown to be possible with a fair degree of success, but complex objects or concepts represent a tougher problem.

Knowledge in context

Physical context

Another facet to the use of artificial intelligence in semantic annotation is that of *context-sensitivity* in mobile devices – devices that can sense their surroundings. Context-sensitivity – brought into the forefront of ubiquitous computing discourse by Dey (2001) – is yet another ‘idea’ technology. The aspiration is a ‘what you see is what you need’ experience, providing services and information without explicit prompting by the user. It is promising, but challenging in practice.

The device-level view of its immediate surroundings is quite different to that of the user. Though it is possible for a device to apply statistical/machine-learning methods in order to learn classifications, or for the developer to hardcode a set of rules defining the characteristics of a given context, it proves difficult to recognise the *same* contexts and features that the device’s owner considers significant. Although a fascinating research field and a technology which is bound to come to the fore with increasing popularity of mobile phones and other wearable devices, context-sensitivity is not a silver bullet. At present, most practical successes in context-sensitivity relate to position, physical state (body monitoring, such as heart rate) and physical activity (walking, running, driving a car).

A camera-phone might tag an image with the location at which it was taken, allow one to browse emails according to the location at which they were written or received, or provide relevant information on a just-in-time basis. Familiar scenarios in the latter research domain include *mediascapes* – contextual availability of information, such as information about pictures in a museum. Identifiers such as Radio Frequency Identification (RFID) tags or barcodes can be used as *keys* to retrieve relevant information, as in landmark projects such as Hewlett-Packard’s Cooltown (HP, 2001) and its descendants, such as BBC Coast and the open Create-A-Scape software³. Mediascapes are typically predefined rather than being dynamically generated from unstructured data, though both are possible. Services such as Panoramio, which interfaces closely with Google Maps, and Flickr, which permits geotagging, can also be used to create a ‘mediascape 2.0’ via user-contributed metadata.



³ http://www.createascape.org.uk/create_a_mediascape/make_mediascape/start_software.html



Several organisations have examined the possibility of location-based image search. These hold the promise of identifying landmarks by taking a photograph and sending it to an appropriate search service, which compares the subject of the image to reference images in order to provide an identification; see for example Davies *et al.* (2005) who describe a mobile tour guide that makes use of a 'point and find' mode of interaction; taking a photograph of an object causes it to be identified (as in Nokia's recent 'Point and Find' prototype). The system then provides information about that object. In practice, subjects often preferred browse modes that provided information about nearby landmarks without requiring the subject to specifically search for each one in order to retrieve information – as with the 'berrypicking' model.

Mobile devices discourage long search sessions due to transmission costs and ergonomic factors, but offer a great deal of information about users' context and information needs, of which commercial interface designers are only now beginning to take advantage. Groups such as GeoVector are exploring commercial use of 'point and click', and recent developments such as Google's free interface software for Google Maps, available for many Nokia devices, are beginning to change this trend.

Social context

When information-retrieval experts talk about context, they seldom think of location or physical variables, but the contexts in which a document was written, in which searches take place, or in which a document is retrieved. The way in which we apply or read symbols is dependent on these variables; both our use of language and search model shift with context of use. Teachers looking for appropriate resources to use in a classroom setting expect different responses to the pupil who types the topic into a search engine in the hope of finding an easy introduction. Context defines method, and is one of the most useful, important yet consistently overloaded and challenging words within and beyond the information retrieval world.

Search engines may provide a service tailored to many aspects of the context of use. Autology, for example, uses information about the user's present task – monitoring a document as it is being written – to tailor search results to the user's personal profile and deliver them 'just-in-time'.

Adaption, personalisation and social search

Searching is hard because there are two degrees of separation between the real world, the way in which things and experiences are represented by our brains or by a computer, and the many ways in which we use language to talk about them. As humans, we have inherited a 'theory of mind', that is, an outstandingly accurate ability to guess at what goes on inside other people's heads. This ability lets us guess at the right way to encode information in

language. We are so good at it that mostly, we are almost completely unaware of it. Computers, on the other hand, are 'mind-blind'. They don't know who we are or where we come from – our socio-cultural background, dialect or perhaps even language. Social network analysis – learning about the structure or society that created the data – is one means to solve this problem.

Computers are limited, simple-minded and cannot apply human standards of interaction. But the user may explicitly inform the computer of his or her preferences by customising their experience. For example, software can be customised to our language preferences, prioritising English results over other hits relevant to the query (although if we are bilingual this is possibly not our intention).

The second solution is the use of subtler forms of personalisation: recommendations based on explicit or implicitly given feedback – collaborative filtering and user profiling. Well-known examples include Amazon's recommendation system. Collaborative filtering looks at similarity; you are an individual and nobody else is quite like you, but there is a group of people who are more similar than others. If you are identified by a system which has learnt something about your preferences and interests as similar to certain other people's, the system can start to test that judgement by showing you the preferences of similar users ('You might like to watch this DVD'). Your reaction to that, if any, helps to narrow down your profile further. Feedback may be implicit (you ignored an option) or explicit (you requested not to see this again).

These methods exploit emergent patterns in society. Though unpredictable as individuals, as society comprises a set of stable structures emergent from individuals, so does our social behaviour and language. We cannot explicitly program the rules that we follow every day, but can generalise across the structures in the aggregate dataset.

Discussion

There is no single 'killer app'. Search engines are blunt instruments designed around the technically feasible. No search engine solves every class of problem with equal facility – probabilistic approaches are by necessity tuned to the average use case. Edge cases (unusual problems) may or may not generate appropriate results.

According to the research, finding the right information is a process, not a single interaction. Search engines cannot answer questions that we have not clearly articulated in our own thoughts, but we can expect some level of support in gaining a better understanding of the topic area, and perhaps can aspire to devices that offer information that we did not yet know we needed. The languages of search and interface will diversify further from the familiar text boxes and submit button, and towards a closer integration with ourselves and the ways in which we choose to work and live.

Mobile and ubiquitous computing hold promise in radically novel and perhaps more intuitive interfaces, such as tangible interfaces that let us explore information about our environment by manipulating elements within that environment directly. Indicating an object might provide us with information about it; building a structure of smart objects might affect our online environment.

It is always dangerous to offer predictions about the future. Successful services are usually those that incrementally enhance our lives – *plus ça change, plus c'est la même chose*. Social networks, mobile phones and the increasing availability of digital media link us more closely together and decrease the importance of physical distance. For this author, the ideal technology is one that performs an analogous task with information, linking search and browse, seamlessly relating different media and resource types, offering new information and starting points for discovery, without annoying the user, applying context-awareness to increase relevance and appropriateness, and rewarding curiosity.

The development of such a system is closely linked to awareness of the hidden structure and habits that make up our day-to-day lives. So next time you hum a tune and wonder about its name, wish you knew the location of the nearest Italian restaurant, or ask yourself why the sky is blue or what that book was you read last month, make a note: you have successfully identified a use case for contextual search.

Conclusion

One message is essential to understanding technology and integrating it into our lives. There is no Google generation and no shortcut to understanding. *People are still people*. With every advance there is a rush of hope and hype; will this technology transcend our limitations? Will the internet democratise society? Can the Semantic Web sweep the complexities of epistemology aside? Will Google teach us information literacy? The answer has always been 'No'. That is still the educator's domain.

The Web offers a wide expanse in which to lose ourselves, but searching and browsing are familiar skills, with familiar pathways for acquisition: attention span, problem-solving, knowledge integration and theory development. As technologies become more familiar to us, we may hope that both teachers and students may spend less time formulating the question and more time making use of the many rich resources on the Web, but computers do not replace motivation and cannot save us the trouble of understanding things ourselves. Indeed, technologies tend to highlight our shortcomings.

Rumours of the death of Education 1.0 have been greatly overstated. Whilst the increasing availability of information of all sorts and the increasingly flexible means by which it may be accessed and filtered are advances that will be welcome to most, this is evolution – not a revolution.

Bibliography

- Alberto Albiol, Torres, L., and Delp, E. J. (2004). 'Face recognition: when audio comes to the rescue of video'. In *Proceedings of the International Workshop on Image Analysis for Multimedia Interactive Services*, Lisboa, Portugal, April 2004.
- Allsopp, John (2006). 'The Big Picture on Microformats'. Digital Web Magazine. http://www.digital-web.com/articles/the_big_picture_on_microformats
- Bates, Marcia. (1989). 'The Design of Browsing and Berrypicking Techniques for the Online Search Interface'. <http://www.gseis.ucla.edu/faculty/bates/berrypicking.html>
- Berners-Lee, T., Hendler, J., and Lassila, O. (2001). 'The Semantic Web'. *Scientific American* 284 (5): 34-43.
- Berry, Michael W. and Browne, Murray (2005). *Understanding Search Engines: Mathematical Modelling and Text Retrieval*, Second Edition. Siam, Philadelphia. Society of industrial and applied mathematics.
- Carneiro, G., Chan, A. B., Moreno, P.J. and Vasconcelos, P. (2007). 'Supervised Learning of Semantic Classes for Image Annotation and Retrieval'. *IEEE Trans. Pattern. Analysis and Machine Intelligence*, March 2007 (Vol. 29, No. 3) pp. 394-410.
- Choo, C. W., Detlor, B. and Turnbull, D. (2000). 'Information Seeking on the Web: An Integrated Model of Browsing and Searching'. *First Monday*, volume 5, number 2 (February 2000). [http://firstmonday.org/issues/issue5_2/choo/index.html]
- Clippingdale, Simon and Fujii, Mahito (2003). 'Face Recognition for Video Indexing: Randomization of Face Templates Improves Robustness to Facial Expression'. *Lecture Notes in Computer Science*, Volume 2849/2003. Springer Berlin / Heidelberg.
- Davies, N., Hesse, A., Dix, A. and Cheverst, K. (2005). 'Understanding the Role of Image Recognition in Mobile Tour Guides', in *Proceedings of the Mobile HCI 05*, Salzburg, Austria.
- Dey, Anind K. (2001). 'Understanding and Using Context', *Personal Ubiquitous Computing* 5(1), 4-7.
- W.I. Grosky, D.V. Sreenath, and F. Fotouhi (2003). 'Emergent Semantics from Users' Browsing Paths,' Proceedings of the First International Conference on Intelligence and Security Informatics, *Lecture Notes in Computer Science*, H. Chen *et al.* (Eds.), Volume 2665, Springer Publishing Company. Berlin, Germany, pp. 355-357.
- Grosky W. I., Sreenath D. V., and Fotouhi, F. (2002) 'Emergent Semantics and the Multimedia Semantic Web'. ACM SIGMOD, Volume 31, Issue 4. Special section on semantic web and data management. Pages: 54 - 58. ISSN:0163-5808.
- Guy, M. and Tonkin, E. (2005) 'Folksonomies: Tidying up tags'. D-Lib 12(1). <http://www.dlib.org/dlib/january06/guy/01guy.html>
- Hewlett-Packard, (2001). Cooltown archive. <http://www.hpl.hp.com/archive/cooltown>
- Hu and B. Liu. (2004). 'Mining and Summarizing Customer Reviews'. In KDD, pages 168-177, Seattle, WA.
- Johnston, Amy E. Boyle, (2007). 'Ray Bradbury: Fahrenheit 451 Misinterpreted. L.A.'s August Pulitzer honoree says it was never about censorship'. *LA Times*, Wednesday, May 30, 2007. <http://www.laweekly.com/news/news/ray-bradbury-fahrenheit-451-misinterpreted/16524>
- Kipp, M., (2006). 'Complimentary or discrete context in online indexing: a comparison of user, creator and intermediary keywords'. *Canadian Journal of Information and Library Science* 30(3). <http://dlist.sir.arizona.edu/1533>
- Lee, Juhnyoung and Goodwin, Richard (2005). 'The Semantic Webscape: A View of the Semantic Web'. WWW 2005. <http://www2005.org/cdrom/docs/p1154.pdf>
- Lee, Jae-Ho, (2005). 'Automatic video management system using face recognition and MPEG-7 visual descriptors'. *ETRI Journal* 27 (6), December 2005
- Marshall, Catherine C. (1997). 'The future of annotation in a digital (paper) world'. Presented at the 35th Annual GSLIS Clinic: Successes and Failures of Digital Libraries

-  McCool, R. (2005). 'Rethinking the semantic Web. Part I'. *IEEE Internet Computing*, Volume 9, Issue 6, Nov.-Dec. 2005 Page(s): 88, 86 – 87 Digital Object Identifier 10.1109/MIC.2005.133.
- Nielsen, J. (2006). 'F-Shaped Pattern for Reading Web Content'. http://www.useit.com/alertbox/reading_pattern.html
- Paolillo, John C., Mercure S., and Wright, Elijah (2005). 'The Social Semantics of LiveJournal FOAF: Structure and Change from 2004 to 2005'. ISWC 2005.
- Perner, P. (2002) Image mining: issues, framework, a generic tool and its application to medical-image diagnosis'. *Engineering Applications of Artificial Intelligence*. Volume 15, Issue 2, April 2002, Pages 205-216.
- Hartmann, K., Büchner, D., Berndt, A., Nürnberger, A. and Lange, C. (2007). 'Interactive data mining & machine learning techniques for musicology'. CIM 07.
- Rieh, Soo Young (2004). 'On the Web at home: Information seeking and Web searching in the home environment'. *JASIST* 55(8): 743-753 (2004). http://www.si.umich.edu/rieh/papers/rieh_jasist2004.pdf
- Shirky, Clay (2003). 'Semantic Web, Syllogism and Worldview'. http://www.shirky.com/writings/semantic_syllogism.html
- Tonkin, Emma., Corrado, Edward M., Moulaison, Heather L., Kipp, Margaret, E. I., Resmini, Andrea., Pfeiffer, Heather D. and Zhang, Qiping (2008). 'Collaborative and Social Tagging Networks'. *Ariadne* 54.
- Widdows, Dominic (2004). 'Geometry and Meaning'. CSLI Lecture Notes No. 172, Center for the Study of Language and Information, Stanford, California.

© Copyright Becta 2008

You may reproduce this material, free of charge, in any format or medium without specific permission, provided you are not reproducing it for financial or material gain. You must reproduce the material accurately and not use it in a misleading context. If you are republishing the material or issuing it to others, you must acknowledge its source, copyright status and date of publication. While great care has been taken to ensure that the information in this publication is accurate at the time of publication, we accept no responsibility for any errors or omissions. Where a specific product is referred to in this publication, no recommendation or endorsement of that product by Becta is intended, nor should it be inferred.

Millburn Hill Road
Science Park
Coventry CV4 7JJ

Tel: 024 7641 6994
Fax: 024 7641 1418

Research email: emtech@becta.org.uk
Main email: becta@becta.org.uk
URL: www.becta.org.uk

