Chapter 7. Key Digital Technologies Underpinning Content & Applications

Yorick Wilks and Matthijs den Besten

(5081 words)

The Internet emerged from efforts to support resource sharing among computer scientists. The technologies being developed to support remote collaboration, resource sharing and other aspects of research in the sciences and humanities are potentially as far reaching for sciences and the humanities as the Internet. This chapter describes in greater depth some of the key technological innovations that lie behind e-Research programmes around the world.

Boxes: Companions, Information Retrieval, Language Processing, RFID, …

Essays:

‘The Semantic Web’ by Wendy Hall and Tim Berners-Lee

‘Embedded Sensor Networks’ by Christine Borgman

Dealing with the Data Deluge

The data deluge is upon us. Several applications, techniques and technologies have been proposed as a way to deal with this data deluge. The purpose of this chapter is to list them.

There are two basic approaches towards the data deluge. One is to try and contain the on-flow of information, the other is to promote it. On one side, there are “guardians of content”, whose aim is to develop representations of the data that are non-ambiguous and noise-free. At the other side, there are “examiners of content”, who rely on the sheer quantity of the data to do the statistical disambiguation for them. Whereas the examiners have developed technologies to acquire and explore more and more detailed data, the guardians have put their energies into more and more sophisticated means to describe these data. We shall see that this tension or complementarity between examiners and guardians plays out at different levels – with regards to the acquisition of data and the creation of content and with regards to the extraction of information and search of content as well. We will look at content and search in turn.

Ubiquitous Computing [14] refers to the trend that we as humans interact no longer with one computer at a time, but rather with a dynamic set of small networked computers, often invisible and embodied in everyday objects in the environment. Alan Kay of Apple calls it ‘Third Paradigm’ computing. Mark Weiser, the father of ubiquitous computing [15], describes it as a, “difficult integration of human factors, computer science, engineering and social sciences”. He states, “Over the next twenty years computers will inhabit the most trivial things: clothes labels (to track washing), coffee cups (to alert cleaning staff to mouldy cups), light switches (to save energy if no one is in the room), and pencils (to digitize everything we draw). In such a world, we must dwell with computers, not just interact with them” and, “We will dwell with these computers, whose presence we will ignore most of the time, and they will provide us with constant clues about our environment, our loved ones, our own past, the objects around us and the world beyond our home.”

Content

The importance of “content” for the Internet, or, at least, the World Wide Web, has long been recognized and, in fact, “content providers” like AOL were in many ways the early stars in that world (Rayport, 1999). Of course, since then things have moved on, but the thrust remains. Most crucially, technologies have emerged that allow a bigger quantity of higher quality content at a lower cost. Turning data into content, emerging technologies allow us to observe the world more intensively and collect large amounts of very detailed data. In addition, the annotation of these observations becomes easier with the day and also less mind-numbing. Finally, more and more powerful techniques are becoming available that help interpret and make sense out of this mass of annotated observations.

The facilities to observe and record the world are growing with the day. Not only can we have cameras at virtually every corner of the street, the low cost of storage makes it feasible for people to record every second of there life.

And images are by no means the only things that can be recorded. Sensors of all types and kinds are being developed and will be placed in areas as remote as the deep ocean or as close as my shallow brain. Two types of technologies greatly enhance the value of all these sensors. They do this by giving the sensors as sense of time and place on the one hand, and a sense of self on the other. Or rather, they make it possible to attach a sensor identifier and a sensor location to the sensor data that are observed. Location can be determined using the infrastructure of GPS and perhaps at some point of Galileo. Object identity can be determined using techniques like RFID.

Awareness of time, location, and provenance of data is great, but usually a great deal of additional information is needed in order to make sense of the data. Luckily, more and more technologies are becoming available that make the elicitation of such information less cumbersome and potentially more fun. First of all, there are efforts to attach meaning to data and make the logical structure of the data explicit as is represented by the Semantic Web. A second class of technologies is less rigorous, but more fun and not necessarily less useful: Mash-ups allow people to make connections between disparate sets of data; tagging allows people to associate labels with the data in structured and unstructured ways; and discussion-fora and blogs allow for an even richer kind of annotation. Finally, experience is growing with ways to induce people to provide increasingly detailed information not only about themselves and their relation to others but also about their understanding of the world at large and interpretation of, say, postcards, in particular (Von Ahn, 2006).

The interpretation of the mass of annotated observations thus generated benefits a lot from the increasingly active role that computer programs take in this environment. Spiders (Mauldin & Leavitt, 1994), programs that crawl sites in search of data to index, have been around a while. More recently, there has been a spate of attempts to develop spider-like agents that are more personalised and search on the basis of the preferences and history the particular individual they intend to serve. Most of these attempts have come to nought. A promising way forward seems to provide users with a richer experience through the development of agents that establish a longer-term companion-like relationship with the users and interact with them in a more natural continuous conversation (Wilks, 2006). Besides, methods like collaborative filtering, using similarity metrics to construct personalised predictions of preferences, and moderation systems play an important role. And “herd computing” (Zittrain, 2007).

Search

The notion of information search is not always easy to separate from that of the content searched for; it depends to some extent on one’s intellectual tradition. Consider the relationship of Information Retrieval (IR), the original search technology, and Artificial Intelligence (AI), whose researchers, according to Karen Spärck Jones, in a remarkable paper (1999), see themselves as what she called “The Guardians of content”. She argued against the mainstream of AI research by which she intended a long tradition of AI work on the representation of knowledge in a computer that has led to developments like the Semantic Web (q.v.). Her view, which can be taken as

the core view of Information Retrieval is that, again in her own words, “One of these [simple, revolutionary IR] ideas is taking words as they stand” (2003), as opposed being wedded, as AI is, to representations, their computational tractability and their explanatory power, over and above the surface words of documents.

IR has been, for fifty years (Berger, Lafferty, 1999), the original search technology for information, and is normally thought of as a technique for selecting the most relevant set of documents from a wider set in response to a user’s query, normally stated as a string of key terms. This idea is not far, in principle, from the everyday democratic use of the internet, where a user types words to Google (on average 2.5 terms per query) to locate information. But that everyday procedure is very different from traditional IR, developed as a technique for library and science professionals, where strings of key terms could be up to 200 long, and where substantial use is now made by search engines like Google of natural language processing (NLP) techniques that the user is not aware of. Those techniques vary enormously in type and scale but what distinguishes them from IR is that make some use of linguistic knowledge, the actual disposition of words in a text, which may be as simple as finding just documents with “Tony Blair” in before finding all those with “Tony” in and all those with “Blair” in, a much larger, and probably less useful, set.

NLP, or language engineering techniques, are now widespread and the most obvious use of them is the machine translation facility available on most search engines. But there remains a lively debate among researchers over whether, in the end, NLP techniques can do anything that sufficiently sophisticated statistical techniques, of IR proper, cannot do. That is the force of the Spärck Jones quotation above: she remained committed to the core IR notion that statistical methods are, in the end, sufficient for good retrieval of information, while those in NLP (and even more in the AI community she criticised) remain equally convinced that some notion of linguistic processing or “understanding of content” is essential for effective information search.

The original search methodology of IR was simply looking for the terms in the query in the document, which is not very successful. The basic strategy for success was to index every document (before search time) with its key or most relevant terms: this used the inverse document frequency (idf) measure of the relevance of terms (Spärck Jones, 1972), the notion that a document is relevant not only because key terms are frequent in it, but because those terms are not frequent in other, non-relevant, documents. All this rested on collecting corpora of non-relevant documents so as to do the indexing. Additional techniques required the use of thesauri, either hand made by experts or created automatically from corpora of related documents, so as to expand sets of relevant terms, so that documents were retrieved that did not exactly match index terms but closely related ones in a thesaurus.

Four further developments rescued IR from a somewhat static condition in the 1980s: first, it was found that one could improve systems by means “relevance feed back”: information from users about which documents found by a system were relevant and which were not. This was the first clear use of

machine learning (ML) in IR, a technique that became essential, see below, to Information Extraction (IE) and Text Mining.

Secondly, is the arrival of the world wide web and the spread of hyperlink algorithms (most famously Brin and Page's, the basis of Google) where relevance is based not on text terms in documents but on pointers to documents. This has led to a culture of web search where queries are normally about two and a half words long (rather than hundreds) so that the ambiguity of terms in short queries s more significant, and the relevance of NLP seems to return. Thirdly, the web has revived Salton's (Salton, 1972) old idea of cross-language IR—retrieving documents in one language by means of a query in another—a technique that he showed to be, surprisingly, as successful as standard (monolingual) IR.

Fourthly and lastly, IR has developed a new mode of analysis in terms of what are called "language models" or "translation models" (Berger and Laferty, 2000). A piece of recent NLP history is highly relevant here: Jelinek, Brown and others at IBM New York began to implement around 1988 a plan of research to import the statistical techniques that had been successful in Automatic Speech Processing (ASR) into NLP and into MT in particular. It was this unfulfilled program of Jelinek, in that it never achieved a success rate in MT of more than 50% of sentences translated correctly, that, more than anything else, began the empiricist wave in NLP that still continues to be its core methodology. Moreover, and this has only recently been noticed, the research metaphors have now reversed, and techniques derived from Jelinek's work are now being introduced into IR under names like "MT approaches to IR" (Berger and Laferty, 1999, and see below) which is precisely a reversal of the direction of influence that KSJ argued for when she claimed that IR and its methods should have more influence on (symbolic) AI and NLP. An extended metaphor is at work here, one where IR is described as MT since it involves the retrieval of one string by means of another. IR classically meant the retrieval of documents by queries, but the string-to-string version notion has now been extended by IR-researchers who have moved on to QA work where they describe an answer as a "translation" of its question (Berger, 2000). On this view questions and answers are like two "languages". In practice, this approach meant taking FAQ questions and their corresponding answers as training pairs.

A quite different and newer search technology is Information Extraction (IE) (Gaizauskas, Wilks, 1997).

IE is an automatic method for locating facts for users in electronic documents (e.g. newspaper articles, news feeds, web pages, transcripts of broadcasts, etc.) and storing them in a data base for processing with techniques like data mining, or with off-the-shelf products like spreadsheets, summarisers and report generators. The historic application scenario for Information Extraction is a company that wants, say, the extraction of all ship sinkings, from public news wires in any language world-wide, and put into a single data base showing ship name, tonnage, date and place of loss etc. Lloyds of London had performed this particular task with human readers of the world's newspapers for a hundred years.

The key notion in IE is that of a “template”: a linguistic pattern, usually a set of attribute-value pairs, with the values being text strings. The templates are normally created manually by experts to capture the structure of the facts sought in a given domain, which IE systems then apply to text corpora with the aid of extraction rules that seek fillers in the corpus, given a set of syntactic, semantic and pragmatic constraints.

IE has already reached the level of success at which Information Retrieval and Machine Translation (on differing measures, of course) have proved commercially viable. By general agreement, the main barrier to wider use and commercialisation of IE is the relative inflexibility of its basic template concept: classic IE relies on the user having an already developed set of templates, as was the case with intelligence analysts in US Defense agencies from whose support the technology was largely developed. The intellectual and practical issue now is how to develop templates, their filler subparts (such as named entities or NEs), the rules for filling them, and associated knowledge structures, as rapidly as possible for new domains and genres. IE as a modern language processing technology was developed largely in the US, but with strong development centres elsewhere (Gaizauskas and Wilks, 1997).

Adaptivity in the MUC development context has meant beating the one-month period in which competing centres adapted their system to new training data sets provided by DARPA; this period therefore provides a benchmark for human-only adaptivity of IE systems. Automating this phase for new domains and genres now constitutes the central problem for the extension and acceptability of IE in the commercial world beyond the needs of the military sponsors who created it. The application of Machine Learning methods to aid the IE task goes back to work on the learning of verb preferences in the Eighties by Grishman and Sterling (1992) and Lehnert (et al., 1992), as well as early work at MITRE on learning to find named expressions (NEs) (Bikel et al., 1997). Many of the developments since then have been a series of extensions to the work of Lehnert and Riloff on Autoslog (Riloff and Lehnert, 1993), the automatic induction of a lexicon for IE.

IE is now widely deployed, on the web and elsewhere, to locate names of certain classes, evens of certain classes, like air crashes in news wires and so on, and it has become to basis of a further NLP technology Question Answering (QA) where the goal is to retrieve from corpora not only relevant facts but the correct answer. IE and QA are quite different from IR in that they rest on linguistic techniques in which the text ceases to be a mere “bag of words” as it traditionally is in IR and, most importantly, higher categories are introduced in the processing—such as capturing all names referring to people or places—which are inherently semantic and involve categories so that words no longer “stand only for themselves” but are grouped into classes with meanings.

Artificial Intelligence (AI), or at least non-Connectionist non-statistical AI, remains committed to the representation of propositions in some more or less logical form. Mainstream IR, as we have seen, is, if not dogmatically anti-representational (as are some statistical and neural net-related areas of AI and language processing), is at least not committed to any notion of representation beyond what is given by a set of index terms, or strings of

index terms along with numbers themselves computed from text that may specify clusters, vectors or other derived structures. However, it is known that many modern search engines for example, do now employ large numbers of people with NLP training and backgrounds and there is no doubt many proprietory engines do now embody techniques (going back to Smeaton and van Rijsbergen, 1988), going far beyond documents indexed as “bags of words”: this is in part due to the ability now to search the full text of documents, rather than the indexing terms. At the simplest level this allows a search engine to distinguish the same terms in different orders, as such classic cases as the undoubted difference of interpretation between:

measurements of models

as opposed to

models of measurement

which might be expected to access different literatures, although the purely lexical content, or retrieval based only on single terms, is the same. In fact they get 363 and 326 hits respectively in Netscape but the first 20 items have no common members.

An extension to syntactic notions in search is that of the use of proposition-like objects as part of document indexing: it could be seen as an attempt to index documents by IE template relations, e.g. if one extracts and filled binary relation templates (X manufactures Y; X employs Y; X is located in Y) so that documents could be indexed by these facts in the hope that much more interesting searches could in principle be conducted (e.g. find all documents which talk about any company which manufactures drug X, where this would be a much more restricted set than all those which mention drug X).

Few notions are new, and the idea of applying semantic analysis to IR in some manner, so as to provide a complex structured (even propositional) index, go back to the earliest days of IR. In the 1960s researchers like Gardin (1965), Gross (1964) and Hutchins (1970) developed complex structures derived from MT, from logic or “text grammar” to aid the process of providing complex contentful indices for documents, entities of the order of magnitude of modern IE templates. Of course, there was no hardware or software to perform searches based on them, though the notion of what we would now call a full text search by such patterns so as to retrieve them go back at least to (Wilks, 1964, 1965) even though no real experiments could be carried out at that time. Gardin’s ideas were not implemented in any form until (Bely et al., 1970), which was also inconclusive. Mauldin (1991), within IR, implemented document search based on case-frame structures applied to queries (ones which cannot be formally distinguished from IE templates), and the indexing of texts by full, or scenario, templates appear in Pietrosanti and Graziadio (1997).

Although this indexing-by-template idea is in some ways an old one, it has not been aired lately, and like so much in this area, has not been conclusively confirmed or refuted as an aid to retrieval. It may be time to revive it again with the aid of new hardware, architectures and techniques. After all,

connectionism/neural nets was only an old idea revived with a new technical twist, and it had a ten year or more run in its latest revival. What seems clear at the moment is that, in the web and Metadata world, there is an urge to revive something along the lines of “get me what I mean, not what I say” (see Jeffrey, 1999). Long-serving IR practitioners will wince at this, but to many it must seem worth a try, since IE does have some measurable and exploitable successes to its name (especially Named Entity finding) and, so the bad syllogism might go, Metadata is data and IE produces data about texts, so IE can produce Metadata.

In IE proper, one can be moderately optimistic that fuller AI techniques using ontologies, knowledge representations and inference, will come to play a stronger role as the basic pattern matching and template element finding is subject to efficient machine learning. One may be moderately optimistic, too, that IE may be the technology vehicle with which old AI goals of adaptive, tuned, lexicons and knowledge bases can be pursued. IE may also be the only technique that will ever provide a substantial and consistent knowledge base from texts, as CYC (Lenat et al., 1986) has failed to do over twenty years. The traditional AI/QA task, now brought within TREC, may yield to a combination of IR and IE methods and it will be a fascinating struggle. The curious tale above, of the use of “translation” with IR and QA work, suggests that terms are very flexible at the moment and it may not be possible to continue to draw the traditional demarcations between IR and these close and merging NLP applications such as IE, MT and QA.

Finally, one must say something about Text Mining (TM, see Kao and Poteet, 2006), a technique that shares with IR a statistical methodology but, being linked directly to the structure of databases, does not have the ability to develop in the way IR has in recent decades by developing hybrid techniques with NLP aspects. Text Mining can be seen as a fusion of two techniques: first, the gathering of information from text by some form of statistical pattern learning and, secondly, the insertion of such structured data into a data base so as to carry out a search for patterns within the structured data, hopefully novel patterns not intuitively observable. These associations can involve associations with other time series information such as stock movements (if the texts are from newspapers). The first of these is more or less coterminous with the IE task, and the second is a more specific form of data mining in general. Users of the term Text Mining often cite tasks such as “Typical text mining tasks include text categorization, text clustering, concept/entity extraction, production of granular taxonomies, sentiment analysis, document summarization” (Kao and Poteet, 2006) and all these are standard and long-standing NLP, IR or IE tasks. Hence the distinctive feature of TM is the statistical search for novel, interesting or relevant relations in such data one extracted.

Challenge or Opportunity?

In this chapter we have seen an array of technologies to deal with the data deluge. We have seen how technologies like ubiquitous computing and sensor networks will exacerbate the existing influx of data and we have seen technologies like the semantic web that promise to funnel it. In addition, we have seen a number of ad-hoc methods for annotation and linkage of data.

Underlying, however, there is a continuing debate as to whether "understanding of content" is essential for effective information search or whether statistical methods will do. Traditionally, practitioners of information retrieval and information extraction would believe the latter while practitioners of natural language processing would opt for the former. More recently, we see a convergence of both traditions where statistical techniques are being adopted in natural language processing and machine translation and search engines now embody techniques that go far beyond indexing documents as "bags of words". Moreover, companions, games, and other applications will allow us to improve the quality and quantity of the data even through the engagement of human brains. It is technologies like these that will ultimately turn the challenge of the data deluge into an unprecedented opportunity.

References

Berger, A. and Lafferty, J. (1999) Information retrieval as statistical translation. SIGIR'99.

Croft, W. and Lafferty, J. (eds.) (2000) Language Modelling for Information Retrieval. Kluwer: Dordrecht.

Gaizauskas, R. and Wilks, Y. (1997) Information Extraction: beyond document retrieval. Journal of Documentation.

Gollins, T. and Sanderson, M. (2001) Improving Cross Language Information Retrieval with triangulated translation. SIGIR'01.

Kao A., Poteet, S. R. (eds.) (2006) Natural Language Processing and Text Mining, Springer, New York.

Salton, G. (1972) A new comparison between conventional indexing (MEDLARS) and automatic text processing (SMART). Journal of the American Society of Information Science, 23(2).

Sparck Jones, K. (1999) Information Retrieval and Artificial Intelligence. Artificial Intelligence Journal, vol. 114.

Sparck Jones, K. (2003) Document Retrieval: shallow data, deep theories, historical reflections and future directions. In Proc. 25th European IR Conference (ECIR03). Lecture Notes in Computer Science. Berlin: Springer. Pp.1-11.

Wilks, Y., Slator, B. and Guthrie, L. (1996) Electric Words: dictionaries, computers and meanings. Cambridge, MA: MIT Press.

[14] http://www.ubicomp2007.org

[15] http://ubiq.com/hypertext/weiser/UbiHome.html