Automatic Link Generation

Ross Wilkinson

CSIRO
Mathematical and Information Sciences
723 Swanston St., Carlton 3053
Melbourne, Australia
Email: Ross.Wilkinson@cmis.csiro.au

Alan F. Smeaton

Dublin City University
School of Computer Applications
Glasnevin
Dublin 9, Ireland
Email: asmeaton@compapp.dcu.ie

In order to access any kind of stored information, one may store it at a specific location, and in the case of electronic information this could be a file name or a Web address. If the location is not known or the amount of information to be accessed is greater than the number of locations that can be remembered, then it is necessary to find the information based on its attributes, its content, or its relationships to other pieces of information whose location is known. In the first two cases, we search, as in information retrieval, while in the latter we navigate, as in hypertext and thus these two areas of hypertext and information retrieval are tightly related [Agosti 1996].

Hypertextual navigation from known locations has some advantages over search. Two of these key advantages are that content creators can provide carefully-defined specific relationships, and that users of the information have a context in which to understand information. However these advantages can be difficult to realise as the size of the information space grows. Some of the problems are:

- The size of the collection may be simply too large to allow for human assigned relationships - a collection of 24,900 articles is difficult to manually cross-reference.
- The collection is sufficiently dynamic that human maintenance costs are too high [Thistlewaite 1997] .
- If the information space is larger than can be authored by a single person, there can be problems with consistency associated with the information space. Ellis et al. noted significant differences in the links manually assigned by different people for the same documents [Ellis 1994].
- When a user enters an information space, they have their own information-seeking context, which develops as they explore the space. Information concerning particular user preferences, what locations have been visited to date, what locations have been visited most frequently, feedback from the user, could all help determine what locations should be suggested for further investigations.

Under these circumstances it makes sense to consider using automatic techniques to determine relationships between pieces of information. Much has been done to address this issue, and for a more comprehensive description, the reader should consult the special issue of Information Processing and Management devoted to the topic [Agosti 1997] . To automatically generate links, we need to first consider what types of relationships are useful. We can

---

have links that relate to a structure associated with the information space such as an overall hierarchy, or we can have links that refer to the semantic closeness of pieces of information, or we can have reference links that point to related though not semantically close pieces of information [Thistlewaite 1997].

Besides the issue of the types of links being created, we should also examine whether the links are static or dynamic. While most links created in the World Wide Web are static, there are good reasons for having links that are determined at the time that a piece of information is being displayed. As has already been mentioned, only in a dynamic hypertext environment can the user's contextual situation, which must be an important consideration in hypertext navigation, be considered. Dynamic link production can also guarantee that no dangling links are displayed, and can allow links to be created from and to material that the link author does not control [Ashman 1997] . The value of creating dynamic links that take context into account demonstrates the importance of open hypertext systems such as Microcosm [Davis 1993] and Hyper-G [Andrews 1995a] .

If source information is already well-described structurally, it is comparatively easy to turn text into hypertext. Early work of Frisse [Frisse 1988] converted a medical handbook using the book's hierachical structure, and with increasing use of SGML and XML to code information it is easy to construct nodes and associated links to represent this information in navigable form [Fahmy 1990] [Fuller 1993].

To create static links between semantically related text, we can simply calculate the similarity between all pairs of information, and then insert links between those that are most similar. This assumes that similarity, as measured by information retrieval techniques, mirrors semantic relatedness, and has been used to good effect. There are many ways of measuring similarity and then determining whether a link should be in place. Many authors have described work on approaches like this and it was particularly important at a time when processing speeds meant that pre-computation was attractive. Furuta et al. describe a comparative study of the quality of links produced [Furuta 1989] . Salton et al. describe building a set of cross-references for an encyclopedia [Salton 1991] and Lelu created links using both similarity and spreading activation [Lelu 1991]. Green introduced the use of lexical chains, exploiting the semantic relatedness of individual words, to determine when links should be used [Green 1998]. Allen demonstrated that there is value in distinguishing between the various sub-types of semantic links, and demonstrated how links associated with these sub-types can be determined and assigned [Allan 1997]. Because links created in this way will vary in quality considerably, it is possible to generate them automatically, but manually vet them [Bernstein 1990] [Chignell 1991].

By integrating a query engine into a hypertext system, it is also possible to create dynamic semantic links. This can be as simple as calculating the similarity between the currently viewed information and all other pieces, and presenting those that are most similar, to explicitly inserting a query as a link [Boy 1991] [Coombs 1990] [Rivlin 1994]. This became attractive as search engines got faster, but fully automatic link generation does have problems. If a similarity threshold is used, there tends to be a very uneven creation of links with huge numbers of links to and from some objects and no paths to others. This effect of this problem can be measured with metrics such as those developed by Botafogo et al. [Botafogo 1992] [Rivlin 1994] which measure the overall structural characteristics of a graph or hypertext in terms of its connectedness or its linearity. In practice, however, such metrics have not been found to be very useful in large or even medium-sized hypertexts as they tend to be suitable for measuring graph topology in a local rather than global setting. In an interesting report on integrating search and hypertext, Tebbutt [Tebbutt 1999] showed that there is value in this integration, but that user's reaction varies widely, and that there are many issues to do with what information is shown, and how, that must be resolved in a successful synthesis. One of the issues that the users raised was the number of links that ought be generated. This raises the larger issue of what is the appropriate structure(s) of a hypertext? Should there be tightly clustered sets of pages created? How should a structure associated with automatically generated links be related to, say, an explicit structure already in place? What models of the information space best support a variety of information discovery activities?

Instead of creating links depending solely on current location, it is also useful to take into account the overall information need of the information seeker. One way of doing this is to dynamically build guided tours, based on an initial description of information need [Guinan 1992]. Calvi and de Bra [Calvi 1997] [Calvi 1998] describe

methods of determining relevant links based on the learning state of the user of a training hypertext. In this case the model of the user is more complicated than is the case in both hypertext (where the location of the user in the information space is known) and in information retrieval (where an explicit query is known). To exploit more fully the power of automatically created links at the time of use, we need more sophisticated user models which take into account not only location, explicit needs, but user history, and long standing user preferences. We may well wish to build discourse models that recognise that exploring an information space is an activity that unfolds.

References

[Agosti 1996] Maristella Agosti and Alan F. Smeaton (editors). Hypertext and Information Retrieval. Kluwer, Boston, 1996.

[Agosti 1997] Maristella Agosti and James Allan. "Methods and tools for the construction of hypertext" in Information Processing and Management, 33(2), 129-271, 1997.

[Allan 1997] James Allan. "Building Hypertext using Information Retrieval" in Information Processing and Management 33(2) 145-159, 1997.

[Andrews 1995a] Keith Andrews, Frank Kappe, and Hermann A. Maurer. "The Hyper-G Network Information System" in Journal of Universal Computer Science, 1(4), 206-220, [Online: http://www.iicm.edu/jucs_1_4/the_hyper_g_network], April 1995.

[Ashman 1997] Helen Ashman, Alejandra Garrido, and Harri Oinas-Kukkonen. "Hand-made and Computed Links, Precomputed and Dynamic Links" in Proceedings of Multimedia '97 (HIM '97), Germany, 191-208, 1997.

[Bernstein 1990] Mark Bernstein. "An Apprentice that Discovers Hypertext Links" in Proceedings of the ACM European Conference on Hypertext '90 (ECHT '90), Versailles, France, 212-223, November 1990.

[Botafogo 1992] Rodrigo A. Botafogo, Ehud Rivlin, and Ben Shneiderman. "Structural Analysis of Hypertexts: Identifying Hierarchies and Useful Metrics" in ACM Transactions on Information Systems, 10(2), 142-180, 1992.

[Boy 1991] Guy A. Boy. "Indexing Hypertext Documents in Context" in Proceedings of ACM Hypertext '91, San Antonio, TX, 51-62, December 1991.

[Calvi 1997] Licia Calvi and Paul de Bra. "Improving the Usability of Hypertext Courseware Through Adaptive Linking" in Proceedings of ACM Hypertext '97, Southampton, UK, 224-225, April 1997.

[Calvi 1998] Licia Calvi and Paul De Bra. "A Flexible Hypertext Courseware on the Web Based on a Dynamic Link Structure" in Interacting with Computers, 10(2), 143-154, 1998.

[Chignell 1991] Mark H. Chignell, Bernd Nordhausen, J. Felix Valdez, and John A. Waterworth. The HEFTI model of text to hypertext conversion. Hypermedia, 3(3):187--205, 1991.

[Coombs 1990] James H. Coombs. "Hypertext, Full Text, and Automatic Linking" in Proceedings of ACM SIGIR '90, Brussels, Belgium, 83-98, September 1990.

[Davis 1993] Hugh C. Davis, Wendy Hall, Adrian Pickering, and Rob J. Wilkins. "Microcosm: An Open Hypermedia System" in Proceedings of ACM INTERCHI'93 Conference on Human Factors in Computing Systems, Formal Video Programme: Hypermedia and Multimedia, 526, 1993.

[Ellis 1994] David Ellis, Jonathan Furner-Hines, and Peter Willett. "On the Measurement of Inter-Linker Consistency and Retrieval Effectiveness in Hypertext Databases" in Proceedings of the 17th Annual International Conference on Research and Development in Information Retrieval, Dublin, Ireland, 51-60, Springer-Verlag, July 1994.

[Fahmy 1990] Eanass Fahmy and David T. Barnard. "Adding Hypertext Links to an Archive of Documents" in The Canadian Journal of Information Science, 15(3), 25-41, 1990.

[Frisse 1988] Mark E. Frisse. "Searching for Information in a Hypertext Medical Handbook" in Communications of the ACM (CACM), 31(7), 880-886, July 1988.

[Fuller 1993] Michael Fuller, Eric Mackie, Ron Sacks-Davis and Ross Wilkinson. "Structured Answers for a Large Structured Document Collection" in Proceedings of ACM SIGIR '93, Pittsburg, PA, 204-213, June 1993.

[Furuta 1989] Richard K. Furuta, Catherine Plaisant, and Ben Shneiderman. "A Spectrum of Automatic Hypertext Constructions" in Hypermedia, 1(2), 179-195, 1989.

[Green 1998] Stephen J. Green. "Automated Link Generation: Can We Do Better than Term Repetition?" in Proceedings of the Seventh International World Wide Web Conference, Brisbane, Australia, 75-84, 1998.

[Guinan 1992] Catherine Guinan and Alan F. Smeaton. "Information Retrieval from Hypertext using Dynamically Planned Guided Tours" in Proceedings of the ACM Conference on Hypertext (ECHT '92), Milano, Italy, 122-130 December 1992.

[Lelu 1991] Alain Lelu. "Automatic Generation of 'Hyper-Paths in Information Retrieval Systems: A Stochastic and an Incremental Algorithms" in Proceedings of ACM SIGIR '91, Chicago, IL, 326-336, October 1991.

[Rivlin 1994] Ehud Rivlin, Rodrigo A. Botafogo, and Ben Shneiderman. "Navigating in Hyperspace: Designing a Structure-based Toolbox" in Communications of the ACM (CACM), 37(2), 87-96, February 1994.

[Salton 1991] Gerald Salton and Chris Buckley. "Automatic Text Structuring and Retrieval - Experiments in Automatic Encyclopedia Searching" in Proceedings of ACM SIGIR '91, Chicago, Illinois, 326-336, October 1991.

[Tanaka 1991] Katsumi Tanaka, N. Nishikawa, S. Hirayama, and Kanayo Nanba. Query pairs as hypertext links. In Proceedings of the Seventh International Conference on Data-Engineering, pages 456--463, Kobe, Japan, 1991. IEEE Computer Science Press.

[Tebbutt 1999] John Tebbutt. "User evaluation of automatically generated semantic hypertext links in a heavily used procedural manual" in Information Processing and Management, 35(1), 1-18, 1999.

[Thistlewaite 1997] Paul Thistlewaite. "Automatic construction and management of large open webs" in Information Processing and Management, 33(2), 161-173, 1997.