GEOGRAPHIC IR SYSTEMS: REQUIREMENTS AND EVALUATION

Bénédicte Bucher[1], Paul Clough[2], Hideo Joho[2], Ross Purves[3] and Awase Khirni Syed[3]

[1]Laboratoire COGIT - Institut Géographique National {benedicte.bucher@ign.fr}
[2]Department of Information Studies, University of Sheffield, UK {p.d.clough;h.joho@sheffield.ac.uk}
[3]Department of Geography, University of Zürich, Switzerland {rsp;sak@geo.unizh.ch}

Abstract

Geographic information retrieval is a new and evolving domain. The development of GIR systems requires the analysis of requirements for such systems and, after systems are implemented their evaluation. This paper describes requirement analysis and evaluation for the SPIRIT system. Methods of user and system-centred evaluation are described and a methodology for building a document collection to facilitate derivation of measures of system performance, together with a new scheme for the evaluation of spatial and thematic relevance with respect to search results are introduced. The paper stresses the importance of developing approaches to evaluate GIR systems which holistically evaluate user interactions as well as measures of system performance.

Introduction

Geographic information retrieval (GIR) is a fast developing area "concerned with providing access to geo-referenced information sources" (Larson, 1996). In recent years information retrieval has become to a large extent synonymous with the retrieval of relevant documents from large collections of unstructured text-based documents stored on the web. In this context, we will define geographic information retrieval more narrowly than Larson, as the retrieval of geographically and thematically relevant documents in response to a query of the form <theme, location> (e.g. Castles, Scotland), where the spatial relationship may either implicitly imply containment, or explicitly be selected from a set of possible topological, proximity and directional options (e.g. inside, near, north of) and where the documents searched are those available on the web. Developments in GIR are driven both by academic enquiry and commercial interests, with for example Google having recently introduced a so-called "Local" search engine based on the integration of data provided by business directories and web documents (http://local.google.co.uk/).

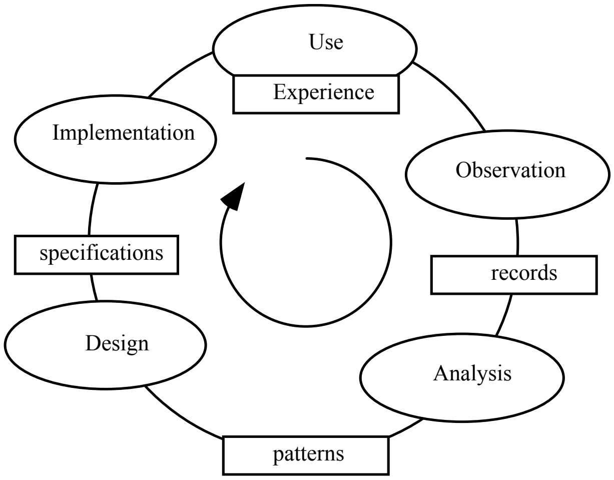


Figure 1: Activities and products of the development process (after Henderson 1991, p262)

The SPIRIT project aims at developing a GIR application that exploits both thematic and locational elements of documents available in the web. It has adopted an iterative development process similar to that shown in Figure 1 which highlights the need for a user-centred design process. These elements can be considered not as start and end points in a development cycle, but rather as key stages to be revisited as software is progressively implemented. Use, observation, analysis and design will be considered in this paper as ***requirements analysis***, resulting in a ***system specification and design*** in the initial phases of a project before any implementation has taken place, and ***evaluation*** when reference to an implemented system is possible.
In this paper we present the results of an initial requirements analysis for a GIR system, SPIRIT (Spatially aware information retrieval on the internet) developed to retrieve documents using both spatial and thematic information from a large collection of unstructured web data for a variety of European locations. We briefly illustrate the implemented system at the time of writing, before discussing how GIR can be evaluated in a general sense and specifically in the SPIRIT project.

Requirements analysis for SPIRT

The basic methodology to analyse the requirements for SPIRIT was two-fold. A set of mock-ups were developed together with scenarios imagined by those responsible for developing various components of SPIRIT. These were presented to potential user groups, and semi-structured interviews were used to collect information about users' views on potential interactions and functionalities of the system. An analysis of existing web-based systems that provide some of the required functionalities was also carried out. For instance we analysed the expression of a place by the user and through browsing maps in applications such as MultiMap (www.multimap.com).Both the analysis of users needs and gaps to fill in the existing applications resulted in a set of functionalities and characteristics that required innovative solutions from the SPIRIT project.

A key requirement for SPIRIT is to retrieve documents with respect to theme, spatial relationship and location. This requirement can be decomposed into several specific functionalities. The first of these is the need to interpret and disambiguate the user's specification of a place name, and provide mechanisms to deal with potentially ambiguous place names (e.g. London, UK vs. London, Ontario) or imprecise regions, such as the south of England (Purves et al., 2005). Having interpreted a query, SPIRIT must submit this query to a search engine which incorporates techniques to deal with the thematic and locational aspects of the query and provide ranked results to a user (Van Kreveld et al., 2004).

Users working through scenarios made clear the importance of displaying results on a map, especially in cases where the spatial relevance of a document with respect to a query is unclear to the user. This typically happens when the user does not have detailed local knowledge of the area under query. Cartographic representation of the query area and the retrieved document's locations allows them to make some assessment of the ***spatial relevance*** of the query results, such as neighbouring or inclusion relationships between a document location and a specific place the user is familiar with. Linking of the ***geographic footprints*** of the retrieved documents with the content of the documents themselves further allows the user to assess the ***thematic relevance*** of the query. The possibility of the user using the interface to redefine their search through, for example selection of the most relevant documents or (re)specification of the query region, was also suggested.

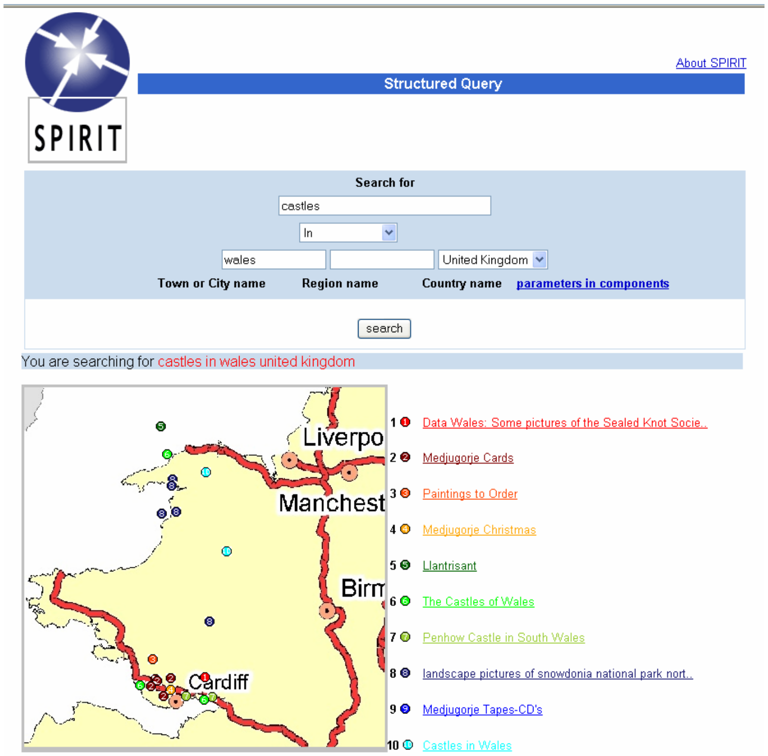
Some users expressed a requirement for additional geographic information in response to their query, for example through locating not only relevant documents, but relevant geographic datasets, e.g. elevation data for a query on "walking in the Alps".

The SPIRIT system

The requirements analysis process described above was used in developing a set of initial considerations for the development of the SPIRIT system. The system itself consists of a number of key components, namely a user interface responsible for query formulation and results display, a search engine and relevance ranking component, responsible for ranking and returning documents based on the triplet of <theme, relationship, location> and a geographical ontology responsible for maintaining information about geographical locations in terms of their semantics and geometry (Jones et al., 2004). However, to the user, the architecture responsible for the response to their query is opaque, and they are presented with results in response to their query through the filter of the user interface (Purves et al., 2005).

The interface itself allows users to formulate their query both textually through a structured query or graphically. In the case of a structured query (Figure 2), the user identifies a location through specifying a place name (e.g. Edinburgh) while for a graphical query a polygon identifying a region of interest is drawn on the map displayed by SPIRIT. A first task in evaluating a SPIRIT is to measure the quality of the results of these individual, differently formulated queries. A second obvious task is to allow intercomparison between the resulting measures to compare the different underlying techniques.

Other SPIRIT components implement several methods, such as for instance the relevance ranking module[1]. The need, and a proposed methodology, to quantitatively and qualitatively evaluate individual components and the differences between such results are discussed in the next section. We put a specific emphasis on proposing evaluation elements that support the comparison of different techniques implemented in SPIRIT prototype and perhaps, in the future, in other GIR applications.



Search for: castles

In

wales | | United Kingdom

Town or City name | Region name | Country name | parameters in components

You are searching for castles in wales united kingdom

1 Data Wales: Some pictures of the Sealed Knot Socie..
2 Medjugorje Cards
3 Paintings to Order
4 Medjugorje Christmas
5 Llantrisant
6 The Castles of Wales
7 Penhow Castle in South Wales
8 landscape pictures of snowdonia national park nort..
9 Medjugorje Tapes-CD's
10 Castles in Wales

Figure 2: SPIRIT's structured query interface and the resulting documents for the query, „Castles in Wales"

Evaluating GIR

GIR is an evolving field. As of yet, few GIR systems based on querying of unstructured text exist, and thus, to our knowledge, limited evaluation has so far been performed of such systems. However, as new systems become available, the need for a framework for the evaluation of GIR systems both in order to measure the performance of individual systems and to compare results between systems will become more pressing.

By contrast, in IR a long tradition of evaluation exists. In most cases, an IR system is used to assist with finding answers (in the form of documents) to a user's information needs[2]. How well a system meets those needs can be evaluated along a number of dimensions. For example, two predominant evaluation strategies have emerged from IR evaluation (Spark Jones and Willett, 1997:167-174): those which are system-focused; and those which are user-centred. In the former, the goal of evaluation is to measure system performance, e.g. to compare and rank different IR systems or components of the same IR system. This approach typically uses an established benchmark to simulate retrieval tasks without requiring involvement of end users. The latter strategy - user-centred evaluation – aims to evaluate IR systems with respect to usability (e.g. assessing the suitability of an IR system interface or some feature of interface design through a task-based user study), involving that the IR system has an interface through which users can interact and upon which observations can be made.

The classic means of measuring the performance of an IR system is based on a standardised evaluation resource called the test collection. This provides the necessary resources and framework in which to assess performance and is typically used to compare different retrieval methods or systems. Approaches to evaluation in IR have long focussed on empirical measures

[1] For example, documents may be ranked based only on their thematic relevance or based on some combined thematic and locational relevance score.

[2] Determining whether or not retrieved documents meet a user's information needs or not is a subjective and controversial topic as the notion of relevance can be defined in many ways (Saracevic, 1975).

of performance which are considered to be objective (e.g. Borlund, 2003, Van Rijsbergen, 1979). Precision and recall are the most commonly quoted formal measures of performance which are produced by such approaches and together are used in the calculation of a wide variety of performance metrics. Both measures are based on systems returning a set of results in descending order of relevance (i.e. the documents assumed to be most relevant are positioned nearer the top). Precision is used to measure how many of the documents returned are relevant; recall measures what proportion of all known relevant documents are retrieved by the system. Further measures include how many relevant are found just in the top 10 results, known as precision at 10. Standard precision and recall can be used only on binary relevance judgement schemes (i.e. relevant vs. non-relevant), but the measures have been generalised to include levels of graded relevance (Kekäläinen and Järvelin, 2002). For example, in the retrieval of structured documents encoded in XML, binary schemes have been found to be insufficiently expressive and graded relevance schemes have been applied (Kazai et al., 2004). Also in image retrieval, more suitable schemes based on ternary judgements have been used (Clough et al., 2005).

The design of a standardised resource for IR evaluation was first proposed almost 40 years ago by Cleverdon (1967) and has since been used in major IR evaluations, such as the Text REtreival Conference or TREC (Vorhees, 2001). In these evaluation campaigns, participants are provided with the same resources and search tasks enabling the comparison of different retrieval systems. Most large-scale evaluation has been system-focused, but these campaigns have also addressed the assessment of user interaction, using data provided by existing IR test collections. A typical IR test collection consists of the following: a set of documents representative of a selected domain, a set of typical user information needs based on the document collection (queries) and a list of which documents are relevant to each query. The generation of such queries and relevance judgements is time consuming. For example, as a sample, the document collection should fairly represent the search domain, the topics should balance between representing realistic user requests and providing controlled queries (Peters, 2001: 1069), and relevance assessments should be as complete as possible (which can be difficult in large document collections).

In recent years the development of systems where the user is allowed more opportunity to interact dynamically during the information seeking process have called into question the use of purely system-driven approaches in evaluating retrieval systems (Borlund, 2003). Rather, a user-oriented approach to evaluation is suggested, where the aim is not to evaluate the quality of a particular document set with respect to a particular query, but rather to measure how the system as a whole facilitates the process of a user seeking and retrieving information.

Evaluating SPIRIT

In evaluating SPIRIT, we consider it appropriate to use both approaches (system and user-oriented) to provide detailed feedback on all aspects of the system. However, the key aspects of user-oriented evaluation do not differ significantly from those of other systems and we set these out only briefly here, before considering in more detail the challenges presented in a system-driven evaluation of SPIRIT.

Borlund (2003) describes one user-oriented approach to evaluating interactive IR systems through the use of a “simulated cover story”, which is a short description of a task. Crucially, relevance is assessed by the user with respect to their task. Interviews may also be carried out to investigate, for example, usability aspects of the system under evaluation. A large literature describing such evaluations of IR systems in particular, and computer software in general exists (e.g. EVALUATION REF) and in the main applies to SPIRIT. Special aspects of SPIRIT which should be considered in a user-oriented evaluation relate to the portrayal, understanding of, and interaction with spatial information (e.g. how users formulate spatial aspects of a query, how they assess the spatial location of a document, how ranking of documents is interpreted, etc…). These aspects are considered by, for instance the geovisualisation and cartography literatures.

The development of measures for use in a system oriented evaluation of SPIRIT requires firstly a test collection, secondly a set of queries for which relevance can be assessed, and thirdly a set of relevance judgments. Building these elements must be done with respect to the spatial nature of the problem and the data and documents specific to SPIRIT.

The first element required is the generation of a document collection. The aims in creating such a collection are firstly, to produce a representative sample of the whole set of documents which may be queried and secondly, to build a collection small enough that it is possible that relevance judgements may be performed for queries on every document in the collection. In the case of the SPIRIT document collection, documents were extracted from a Terabyte web collection (Joho and Sanderson, 2004) to build this document collection.

The document collection must contain documents relevant to queries that are likely to be tested on it and in GIR this means consideration must be given to spatial relationships. For example, if we are going to assess a query such as “pubs near Glencoe” we must have documents from locations near Glencoe within our document set. The technique used to build the SPIRIT document set was as follows:

- Choose a set of queries
- Perform a search for the query set from the Terabyte collection
- Record the top N documents from each query
- Generate a document pool from the previous stage and remove any duplicates
- Fetch the full-texts for the unique documents
- Index the fetched texts

We experimented with two methods for generating a set of queries. In the first approach we used the names of the 200 largest (by population) towns and cities in the UK. For each of the 200 names, the top 50 documents were recorded (although several names did not return a full set of 50 documents) generating a collection of 9,010 documents (approximately 85MB of text). In the second approach we based the set of queries on topics thought to be representative of user needs for a GIR system such as SPIRIT (e.g. “Arts festivals in Edinburgh”). In addition, we also used the location without concept as a query (e.g. “Edinburgh”) to provide coverage for different topics. The main focus, again, was the UK, although queries for Montreux in Switzerland we are also included.

To provide data to test spatial operators like “near”, we also used queries based on locations manually judged to be near to a given location (e.g. “Caerphilly” as a place near to “Cardiff”). For each query, the top 1000 documents were used (where applicable) to generate the collection. In total, 21,094 documents (approximately 1GB of text) were retrieved and included in this collection. In both cases, text-based searching was used to perform a search from the TB collection which involved no disambiguation of ambiguous place names, e.g. documents about “Lancaster” will include those for Lancaster Pennsylvania (USA) and Lancaster, Lancashire (UK). However, the main focus of both collections is predominantly the UK.

The generation of queries for testing of a GIR system requires consideration of a number of elements specific to searches of the form <theme, relationship, location>. Furthermore, the nature of queries should test the capabilities of GIR which are not available in standard IR systems and where spatial awareness is required:

- The name of a location is also a more commonly used word (e.g. *Battle*)
- The location name is ambiguous (e.g. *London*, Ontario or *London*, England)
- The location name refers to an imprecise region (e.g. The *south of France*)
- The information required is searched for at a location unlikely to be referred to in web documents (e.g. *blogs near Vélieux*)
- The theme of the query includes a (spatially) irrelevant location (e.g. *Elgin Marbles in Athens* – Elgin is also a town in the north of Scotland)
- The query includes a non-containment based spatial relationship (e.g. *golf courses north of Aberdeen*)

Having developed a set of queries the final step before deriving metrics is to measure relevance for a particular query. In the case of SPIRIT it was decided to develop a scheme to measure both the thematic and spatial relevance. The initial scheme is shown in Table 1.

Thematic relevance

1. A document which contains relevant information about the concept queried AND on its own allows you to form a judgment about the document (i.e. requires no external knowledge).
2. A document is relevant, since it points to a resource MENTIONING the concept, but you must consult further pages referenced by the document to perform a judgment.
3. A document does not provide information about the concept provided.

Spatial relevance

1. A document refers to a location that is/near the query location AND you think that the location in the document has sufficient detail for you to find it on a local map of the area
2. A document refers to a location that is in/near the query location BUT you think that there is insufficient information for you to find that location on a local map of the area
3. A document does not fall within the query location

Table 1: Initial scheme used by assessors to judge spatial and thematic relevance to a query

In this scheme, we structure a relevance judgement in two components : thematic relevance and spatial relevance. We also propose a ternary scheme to differentiate between levels of relevance. Importantly, spatial relevance also has some theme dependence, since the local map to some extent corresponds the spatial granularity of the theme. For instance, for the query "churches in *" the spatial relevance will be high (1) if a full address is given within a city, but still high if only a village name is given for a small village. On the other hand, if only a town or a county name are given we would consider the spatial relevance to be (2).

Since this scheme was new it was decided to test its usability. A total of 10 documents were retrieved for each of 5 queries (Table 2), and 11 subjects judged each document for its spatial and thematic relevance (giving a total of 1100 judgements). The scheme was illustrated through the use of examples to the relevance assessors. Spatial relevance was intended to address issues of granularity that is the detail, as well as the relevance, of spatial information in a document.

1. Caving in Derbyshire (UK)
2. Castles in Wales (UK)
3. Skiing near Glencoe (UK)
4. Art festivals in Edinburgh (UK)
5. Music in Montreux (Switzerland)

Table 2: Queries used in testing relevance scheme

A simple questionnaire was devised to investigate how well users understood and could apply these schemes. In general users found the schemes both easy to understand and suitable, but they were less confident in applying the spatial scheme. We believe this difficulty results when a user is not familiar with an area, and clues to spatial relevance are not simply the place name in the query. For instance, a ski area near Glencoe exists but is called the „White Corries“. A Swiss or French relevance assessor is unlikely to have this detailed level of local knowledge and thus finds it hard to judge the spatial relevance of the page returned. However, if the user is presented with results and a map showing the location of the website, then the spatial relevance of the document is confirmed. This has two implications for relevance assessments in GIR. Firstly, relevance assessment of documents with respect to spatial aspects is best done by assessors with local knowledge of the area under query. Secondly, care should be taken in measuring spatial relevance since users unfamiliar with an area are likely to classify spatial relevance using contextual clues (e.g. locational query words in the document).

We further assessed agreement between assessors for spatial and thematic relevance using a multirater Cohen's Kappa for all 50 documents across the 5 topics. For both thematic and spatial relevance the agreement between annotators was found to be significant across the entire set of tests at the 95% level. However, many assessors also commented on this measure of thematic relevance being unnecessarily cumbersome and expressed a preference for a binary scheme. Thus, it is likely that future evaluation of SPIRIT results will use a standard binary scheme for thematic relevance and a ternary scheme for spatial relevance. Furthermore, more research is required with a larger number of annotators to test the effectiveness of these relevance schemes, particularly for spatial relevance.

Summary and further work

This paper has described the process of requirements analysis and evaluation for a GIR system, SPIRIT. Requirements analysis was carried out through the use of mock-ups and scenarios, and the resulting system has now been implemented.

Evaluation of GIR systems in general and SPIRIT in particular requires that existing evaluation techniques are adapted, and new schemes developed to measure, for example relevance. Furthermore, the high level of interactivity of GIR systems requires that more consideration is given to the user of techniques for evaluating systems with high levels of interactivity. In order to calculate measure to allow assessment of system performance and comparitative studies a document collection has been built on which spatial and thematic relevance assessments can be performed for a variety of queries.

Initial assessments suggest that inter-annotator agreement for a ternary spatial relevance scheme is good, although care is required in the selection of assessors with sufficient knowledge of the region under investigation. In A Corũna we will present results of the evaluation of the final prototype of the SPIRIT system. A further formalisation of relevance in the field of GIR should consider modelling spatial activities and the affordances of geographical information, where affordances of a location are what it offers through its environment (Jordan et al., 1998).

Acknowledgements

This research is supported by the EU-IST Project No. IST-2001-35047 (SPIRIT) and the Swiss BBW (01.0501).

References

Borlund, P. 2003. The IIR evaluation model: a framework for evaluation of interactive information retrieval systems. In: *Information Research*, vol. 8, no. 3, paper no. 152.

Cleverdon, C. The Cranfield test on index language devices: In Sparck Jones, K. & Willett, P. eds. (1997) *Readings in Information Retrieval*. San Francisco: Morgan Kaufmann. Pp. 47-59.

Clough, P., Mueller, H. and Sanderson, M. (2005), The CLEF 2004 Cross Language Image Retrieval Track, *In Multilingual Information Access for Text, Speech and Images: Results of the Fifth CLEF Evaluation Campaign,* Eds (Peters, C., Clough, P., Gonzalo, J., Jones, G., Kluck, M. and Magnini, B.), Lecture Notes in Computer Science (LNCS), Springer, Heidelberg, Germany (in print).

Henderson A. 1991. A Development Perspective on Interface Design and Theory in J.M. Carroll (Ed): *Designing Interaction: Psychology at the Human Computer Interface*. Cambridge, Cambridge University Press: 254-268.

Joho, H., and Sanderson, M. 2004. The SPIRIT Collection: an overview of a large web collection. *SIGIR Forum*, 38(2).

Jones, C.B., Abdelmoty, A.I., Finch, D., Fu, G. & Vaid, S. 2004. The SPIRIT Spatial Search Engine: Architecture, Ontologies and Spatial Indexing. In *Proceedings of the 3rd International Conference on Geographic Information Science* (GIScience 2004), Maryland, USA. LNCS.

Jordan, T. M. Raubal, B. Gartrell, and M. Egenhofer. 1998. An Affordance-Based Model of Place in GIS. in: T. Poiker and N. Chrisman (Eds.), 8th Int. Symposium on Spatial Data Handling, SDH'98, Vancouver, Canada, pp. 98-109.

Kazai, G., Lalmas, M., De Vries, P. 2004. The overlap problem in content-oriented XML retrieval evaluation. SIGIR 2004: 72-79

Kekäläinen, J. & Järvelin, K. 2002. Using graded relevance assessments in IR evaluation. *Journal of the American Society for Information Science and Technology 53*(13), 1120-1129.

Larson, R.R. 1996. Geographic Information Retrieval and Spatial Browsing. In *GIS and Libraries: Patrons, Maps and Spatial Information*, Linda Smith and Myke Gluck, Eds., University of Illinois.

Purves, R.S., Clough, P. and Joho, H. 2005. Identifying imprecise regions for geographic information retrieval using the web. In *Proceedings of GISRUK*, 2005.

Saracevic, T. 1975. Relevance: a review of and a framework for the thinking on the topic. *Journal of the American Society for Information Science*, vol. 26: 321-343

Sparck Jones, K. & Willett, P. eds. 1997. *Readings in Information Retrieval.* San Francisco: Morgan Kaufmann.

Van Kreveld, M., Reinbacher, I., Arampatzis, A. and Van Zwol, R. 2005. Multi-Dimensional Scattered Ranking Methods for Geographic Information Retrieval. *Geoinformatica*, 9,1, 61-84.

Van Rijsbergen, C. J. Information Retrieval, Butterworth-Heinemann, Newton, MA, 1979 (http://www.dcs.gla.ac.uk/Keith/Preface.html)

Voorhees, E.M. 2001. Overview of TREC 2001. In Proceedings of TREC 2001, NIST.