

**Information Retrieval  
or  
'Experimental Retrieval' according to White & McCain<sup>1</sup>**

**Group paper and presentation**

**By**

**Melanie Swain**

**Julie Valvo**

**Stephanie Bond**

**Denise A. Wallace**

**April 26, 2008**

## Information Retrieval

or

'Experimental Retrieval' according to White & McCain<sup>1</sup>

Information retrieval is a branch of information and /or computer science that concerns itself with "the retrieval of information (not data) from a collection of written documents. The retrieved documents aim at satisfying a user information need usually expressed in natural language."<sup>2</sup> The term 'experimental retrieval' is not the common moniker used in the profession and is mainly found in references to the White and McCain article on co-citation analysis where it is defined as focusing "on the design and evaluation of document retrieval systems...working with content neutral indexing theory, thought experiments, or document testbeds."<sup>3</sup>

"Document retrieval systems" are generally called information retrieval systems and are defined by Harter as "a device interposed between a potential end-user of an information collection and the information itself. For a given information program, the purpose of the system is to capture wanted items and filter out unwanted items from the information collection"(p.2).<sup>4</sup>

The nature of distinguishing between 'wanted items' and 'unwanted items' is articulated using the terms recall and precision. Recall is defined as "an information retrieval performance measure that quantifies the fraction of known relevant documents which were effectively retrieved."<sup>5</sup> Precision is defined as "an information retrieval performance measure that quantifies the fraction of retrieved documents which are known to be relevant."<sup>6</sup> These relationships are mathematically expressed as:

$$\text{Recall} = \frac{\text{Number of relevant documents retrieved}}{\text{Total number of relevant documents in the file}}$$

$$\text{Precision} = \frac{\text{Number of relevant documents retrieved}}{\text{Total number of documents retrieved in the file}}$$

Defining and measuring what is determined 'relevant' is at the core of information retrieval studies. Rubín states that "There are at least two aspects of relevance: relevant to the user and relevant to the topic. In the former, it is clear that the user defines the context for relevance. An item retrieved from an information system is relevant if the user believes that it helps to meet his or her information need. In the latter case, an item is relevant if it can be shown that it is about the subject, regardless of a given user. Relevance forms the basis of much evaluation of information systems. Systems that retrieve relevant items and avoid the retrieval of irrelevant items naturally are considered more effective."<sup>7</sup>

According to Persson, information retrieval studies can be divided into two parts: researchers working in algorithms and researchers working with the user-system relationship.<sup>8</sup> The algorithmic researchers make up half of the heaviest cited 'experimental retrievalists' in the White and McCain article.<sup>9</sup> The user centered researchers comprise the other half of the list and share interests with the 'online retrievalist' but are gaining in popularity as user-system relationships are at the front of current research topics outside of the scope of the White and McCain research period.<sup>10</sup>

This report is an attempt to present a brief synopsis of the major players in the history of information retrieval and will focus on the contributions of Gerard Salton, Cyril W. Cleverdon, Stephen E. Robertson, Bruce Croft, Martin Dillon, Karen Spärck Jones, and C.J. van Rijsbergen.

The contributions of Gerard Salton and C.J. van Rijsbergen will be covered by Denise A. Wallace, Cyril W. Cleverdon by Stephanie Bond (and Denise A. Wallace), Karen Spärck Jones and Martin Dillon by Julie Valvo, and Stephen E. Robertson and Bruce Croft by Melanie Swain.



### **Cyril W. Cleverdon (1914-1997)**

Cyril was an English information scientist and a librarian at the Cranfield College of Aeronautics (u.K., later to become Cranfield Institute of Technology). In 1946, Cyril began work on what later became known as the “Cranfield Experiments”.<sup>11</sup> “This work was one of the most important contributions that shaped the field of information science in the 1950’s and 60’s.”<sup>12</sup>

In working with the “Cranfield Experiments”, “Cyril developed the mathematics of ‘recall’ and ‘precision’ as measures of Information Retrieval systems, and built the first test collections for measuring them.”<sup>13</sup> In addition to the uses of ‘recall’ and ‘precision’, Cyril and his colleagues Mills and Keen also included four other factors in determining effectiveness of an information retrieval system. Those four factors include: coverage of the collection, the time lag from query to response, the presentation of output, and the effort of the user.<sup>14</sup>

Cyril W. Cleverdon was awarded the Gerard Salton Award in 1991 for his work with the “Cranfield Experiments”.<sup>15</sup>



### **Gerard Salton (1927-1995)**

Gerald Salton was born in Nuremberg, Germany in 1927; during WWII he was forced to flee Germany. Salton arrived in the United States in 1947 and attended Brooklyn College where he completed his degree in mathematics in 1950. After completing his master’s degree he went on to gain his Ph.D in Applied Mathematics at Harvard. Salton taught for several years at Harvard before joining the faculty at Cornell University. At Cornell, Salton co-founded the Department of Computer Science and dedicated his research to information retrieval and computer science. He remained at Cornell University until his death in 1995.<sup>16</sup>

Gerard Salton is considered the father of information retrieval and is credited with being “the man most responsible for the establishment, survival, and recognition of Information Retrieval (IR).”<sup>17</sup>

Salton’s contributions to the field of IR are great. He is known for his work with the vector space model, term weighting, relevance feedback, clustering, extended Boolean retrieval, term discrimination value, dictionary construction, term dependency, text understanding and structuring, passage retrieval, and most notably, automatic text processing using SMART (Salton’s Magic Retriever of Text or System for the Manipulation and Retrieval of Text).

SMART started at Harvard but came to fruition during Salton's years at Cornell. It is "a powerful experimental retrieval" system that has become "the standard upon which modern information retrieval systems are based on."<sup>18</sup>

The Special Interest Group on Information Retrieval (SIGIS) presents the Gerard Salton Award every three years for those who have made "...significant sustained and continuing contributions to research in information retrieval." Gerard Salton was the first recipient of the award in 1983.<sup>19</sup>

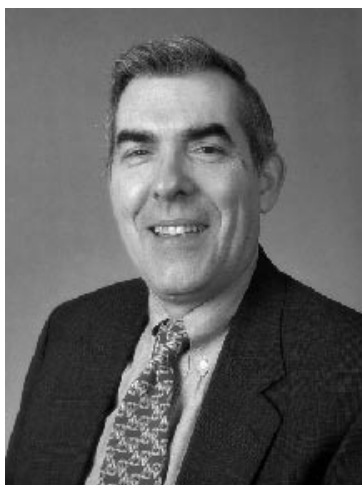


#### **Karen Spärck Jones (1935-2007)**

Born in Huddersfield, Yorkshire, England, Jones studied history and later philosophy at Girton College, Cambridge until 1956. After graduating, for a brief amount of time, she became a school teacher. She went on to start her research career, in the late '50s, working on her doctorate at the Cambridge Language Research Unit.<sup>20</sup> In 1964, she wrote her PhD thesis, "Synonymy and Semantic Classification" exploring combined statistical and symbolic techniques in NLP, which was later recognized to be far ahead of its time. Twenty years after she wrote the thesis, she was encouraged to believe that the work was still relevant and was urged to publish it.<sup>21</sup>

Jones worked at Cambridge's Computer Laboratory as the Professor of Computers and Information and retired in 2002. She continued to work in the Laboratory until shortly before her death. Her most recent work was on information and language system evaluation, user and agent modeling, database query, and document retrieval including speech applications.<sup>22</sup>

Jones is considered one of the pioneers in information retrieval and natural language processing. Her most important contributions to the field are thought to be, at this point, the concept of inverse document frequency (IDF), which is used in most search engines today. She published more than 200 significant papers and nine books.<sup>23</sup> Jones won many rewards for her works including the Gerard Salton Award, ASIS&T Award of Merit, the ACL Lifetime Achievement Award, the BCS Lovelace Medal, and the ACM-AAAI Allen Newell Award.<sup>24</sup>



#### **Dr. Martin Dillon (1938-)**

Dr. Dillon graduated from Canisus College in 1961 and earned a doctoral degree in English from the State University of New York at Buffalo in 1967. From 1969 until 1985, he served as professor, at the University of North Carolina at Chapel Hill, in the school of Information and Library Science. With his teachings and research specialties, he focused on information retrieval and library automation.<sup>25</sup>

He was the visiting distinguished scholar in OCLC's Office of Research in 1985, and by 1986, he had become its director. Dillon led a team of 30 staff members researching and working toward achieving OCLC's goal of improving information access. In 1993, he became the director of OCLC's Library Resources Management Division, which managed the Cataloging and Resource Sharing services.<sup>26</sup>

As a prolific writer, Dillon has published numerous articles within the field of library and information science including literature on topics in information retrieval and automation. He currently serves as the consulting editor for Scarecrow Press and also operates a private technology consulting firm.<sup>27</sup>



### **Stephen E. Robertson (1946-)**

With a strong background in mathematics, Robertson earned his PhD. in Information Studies at University College in London. He currently works as a researcher at the Microsoft Research Laboratory which is located in Cambridge, UK. From 1978 to 1998, Robertson was on the full-time faculty at the City University. While at City University, Robertson served as head of department from 1988 to 1996, and he was instrumental in the development and instrumentation of the Centre for Interactive Systems Research in the Department at City University.<sup>28</sup>

Stephen E. Robertson and Karen Spärck Jones co-wrote an article in 1976 on the probabilistic theory of relevance weighting. This article established him in the field and led to the development of BM25 function which is used in document scoring and term weighting. Robertson's main research interests are in theories and models for information retrieval. More specifically, he is involved with research of probabilistic models along with the design and evaluation of Information Retrieval systems.<sup>29</sup>

Robertson has served on Editorial boards like the Journal of the American Society for Information Science and Technology, and he has received several awards including the Tony Kent Strix award (1998) and the Gerard Salton award (2000).



### **Cornelis Joost "Keith" van Rijsbergen (1943-)**

C.J. van Rijsbergen was born in Rotterdam, Holland in 1943. His formal education spans four continents culminating in a degree in mathematics from University of Western Australia, a graduate degree in computer science and a Ph.D in computer science from Cambridge University in 1972.<sup>30</sup> C.J. has taught information retrieval and artificial intelligence at Monash University, held a Royal Society Information Research Fellowship at Cambridge, and was appointed chair of computer science at the University College of Dublin. He is currently a professor of computer science at the University of Glasgow, leader of the Glasgow Information Retrieval Group and chairman of the Scientific Board of the Information Retrieval Facility.<sup>31</sup>

C.J. van Rijsbergen is considered one of the founders of modern information retrieval; his work on the use of hierarchic clustering with N. Jardine introduced more effective and efficient retrieval when working with large document collections, and he is the author of the seminal monograph *Information Retrieval*, Butterworths 1979 and of the textbook *The Geometry of Information Retrieval*, CUP 2004.<sup>32</sup>

*Information Retrieval*, Butterworths 1979 presented C.J.'s ideas concerning the use of probabilistic modeling and put forth an approach to understanding information retrieval through a non-linear history

of information retrieval. In addition he divided information retrieval down into three areas of study: content analysis, information structures, and evaluation; and noted many of the problems and possible solutions associated with modern information retrieval.<sup>33</sup>

C.J. Rijsbergen was awarded the Tony Kent Strix award in 2004 and the Gerard Salton Award in 2006; his current research is based on context-sensitive information retrieval based on quantum theory with applications to cross-media searching and structured document access.<sup>34</sup>



### **W. Bruce Croft**

Bruce Croft is currently the Director of the Center for Intelligent Information Retrieval. Since 1979, Croft has served on staff as a faculty member at the University of Massachusetts, Amherst, in the Department of Computer Science.

In 1973, Croft received a B.S. and in 1974, he completed his M.S. in Computer Science from Monash University in Melbourne, Australia. Croft continued his education at the University of Cambridge, England, where he received his Ph.D.

in Computer Science. His research interests are primarily in the Information Retrieval genre with specialties in areas such as web search engines, cross-lingual retrieval, and information retrieval models. Recently, Croft has become involved in the Digital Library field which is natural relationship between his information retrieval research interests and his computer science background.<sup>35</sup>

Croft is a member of many organizations, and he has served on several Editorial Boards like the ACM Transactions of Information Systems (1995-2002). Bruce Croft has received many awards. He was the Fellow of ACM (1997), the Research Award from the American Society for Information Science and Technology (2000), and the Gerard Salton Award (2003).

### **Information Retrieval Citations**

1. Howard White and Katherine W. McCain, "Visualizing a Discipline: An Author Co-Citation Analysis of Information Science, 1972-1995," *Journal of the American Society for Information Science*, 49, no. 4 (April 1998): 336.
2. Ricardo Baeza-Yates and Berthier Ribeiro-Neto, *Modern Information Retrieval*, 1<sup>st</sup> ed.(n.p.:Addison Wesley, 1999), under "Glossary," <http://people.ischool.berkeley.edu/~hearst/irbook/glossary.html> (accessed April 23, 2008).
3. Howard White and Katherine W. McCain, "Visualizing a Discipline: An Author Co-Citation Analysis of Information Science, 1972-1995," *Journal of the American Society for Information Science*, 49, no. 4 (April 1998): 336.
4. Stephen Harter, *Online Information Retrieval*, (Orlando, Fla.: Academic Press, 1986), quoted in Richard Rubin, *Foundations of Library and Information Science*, 2ed.(New York, NY: Neal-Schuman Publishers, Inc., 2004): 48.

5. Ricardo Baeza-Yates and Berthier Ribeiro-Neto, *Modern Information Retrieval*, 1<sup>st</sup> ed.(n.p.:Addison Wesley, 1999), under "Glossary," <http://people.ischool.berkeley.edu/~hearst/irbook/glossary.html> (accessed April 23, 2008).
6. Ricardo Baeza-Yates and Berthier Ribeiro-Neto, *Modern Information Retrieval*, 1<sup>st</sup> ed.(n.p.:Addison Wesley, 1999), under "Glossary," <http://people.ischool.berkeley.edu/~hearst/irbook/glossary.html> (accessed April 23, 2008).
7. Richard Rubin, *Foundations of Library and Information Science*, 2ed.(New York, NY: Neal-Schuman Publishers, Inc., 2004): 48.
8. Olle Persson, "The Intellectual Base and Research Fronts of JASIS 1986-1990," *Journal of the American Society for Information Science*, 45, no. 1 (January 1994): 31-38, quoted in Howard White and Katherine W. McCain, "Visualizing a Discipline: An Author Co-Citation Analysis of Information Science, 1972-1995," *Journal of the American Society for Information Science*, 49, no. 4 (April 1998): 346.
9. Howard White and Katherine W. McCain, "Visualizing a Discipline: An Author Co-Citation Analysis of Information Science, 1972-1995," *Journal of the American Society for Information Science*, 49, no. 4 (April 1998): 334.
10. Dangzhi Zhao and Andreas Strotmann, "Information Science during the First Decade of the Web: An Enriched Author Co-citation Analysis," *Journal of the American Society for Information Science and Technology*, 59 (2008): 916-937, [http://www.ualberta.ca/~dzhao/Zhao\\_Dangzhi\\_JASIST08\\_preprint.pdf](http://www.ualberta.ca/~dzhao/Zhao_Dangzhi_JASIST08_preprint.pdf) (accessed April 23, 2008).
11. Birger Hjørland, "Cleverdon, Cyril William (1914-1997)," <http://www.db.dk/bh/core%20concepts%20in%20lis/articles%20a-z/cleverdon.htm> (accessed April 23, 2008).
12. Josh Reiss and Mark Sandler, "Beyond Recall and Precision: A Full Framework for MIR System Evaluation," (London, 2002) <http://www.elec.qmul.ac.uk/people/josh/documents/ReissSandler-Ismir-2002-BeyondRecallandPrecision.pdf> (accessed April 23, 2008).
13. Michael Lesk, "The Seven Ages of Information Retrieval," International Federation of Library Associations and Institutions, Activities and Services, UDT Occasional Paper #5, (June 17, 1995) <http://www.ifla.org/VI/5/op/udtop5/udtop5.htm> (accessed April 23, 2008).
14. Josh Reiss and Mark Sandler, "Beyond Recall and Precision: A Full Framework for MIR System Evaluation," (London, 2002) <http://www.elec.qmul.ac.uk/people/josh/documents/ReissSandler-Ismir-2002-BeyondRecallandPrecision.pdf> (accessed April 23, 2008).
15. "Awards", Special Interest Group on Information Retrieval, Gerard Salton Award, <http://www.sigir.org/awards/awards.html#salton> (accessed April 23, 2008).
16. Karla Consroe, "In Memoriam," Cornell University, <http://www.cs.cornell.edu/Info/Department/Annual96/Beginning/salton.html> (accessed April 23, 2008).

17. *ibid.*

18. *ibid.*

19. "Awards", Special Interest Group on Information Retrieval, Gerard Salton Award, <http://www.sigir.org/awards/awards.html#salton> (accessed April 23, 2008).

20. "Karen Spärck Jones: Researcher whose work on information retrieval underpins modern search technologies on the world wide web," Obituaries, *The Times*, June 22, 2007, TimesOnline, <http://www.timesonline.co.uk/tol/comment/obituaries/article1968942.ece> (accessed March 26, 2008).

21. "Karen Spärck Jones (1935-2007)," University of Cambridge, Computer Laboratory, Obituaries, <http://www.cl.cam.ac.uk/misc/obituaries/sparck-jones/> (accessed March 26, 2008).

22. *ibid*

23. "Karen Spärck Jones: Researcher whose work on information retrieval underpins modern search technologies on the world wide web," Obituaries, *The Times*, June 22, 2007, TimesOnline, <http://www.timesonline.co.uk/tol/comment/obituaries/article1968942.ece> (accessed March 26, 2008).

24. Rowe, Josiah, "Karen Spärck Jones," Wikipedia, [http://en.wikipedia.org/wiki/Karen\\_Sp%C3%A4rck\\_Jones](http://en.wikipedia.org/wiki/Karen_Sp%C3%A4rck_Jones) (accessed March 26, 2008).

25. Dean, Nita, "Martin Dillon to Head New OCLC Institute," OCLC Institute (January 14, 1997) <http://www5.oclc.org/downloads/press/releases/1997/970114c.htm> (accessed March 26, 2008).

26. "Bicentennial Conference on Bibliographic Control for the New Millennium," Library of Congress (June 27, 2000) <http://www.loc.gov/catdir/bibcontrol/dillon.html> (accessed March 26, 2008).

27. Wells, Jenny, "Annual Lazerow Lecture to be Held April 7," University of Kentucky, (March 26, 2008) [http://news.uky.edu/news/display\\_article.php?artid=3314](http://news.uky.edu/news/display_article.php?artid=3314) (accessed April 12, 2008).

28. "Stephen E. Robertson: Non-Stipendiary Fellow," Girton College, University of Cambridge, <http://girton.etianen.com/fellows-and-staff/non-stipendiary-fellows/s-robertson/> (accessed April 23, 2008).

29. "Stephen E. Robertson Researcher," Microsoft Research (February, 2005) <http://research.microsoft.com/users/robertson/> (accessed April 23, 2008).

30. "C.J. 'Keith' van Rijsbergen," Department of Computer Science, University of Glasgow, <http://www.dcs.gla.ac.uk/~keith/> (accessed April 23, 2008).

31. "C.J. van Rijsbergen," Wikipedia, [http://en.wikipedia.org/wiki/C.\\_J.\\_van\\_Rijsbergen](http://en.wikipedia.org/wiki/C._J._van_Rijsbergen) (accessed April 23, 2008).

32. N. Jardine and C. J. van Rijsbergen, "The Use of Hierarchic Clustering in Information Retrieval," *Information Storage and Retrieval*, 7, 5, (December 1971): 217-240.

33. C. J. Rijsbergen, *Information Retrieval*, 2 ed. (Department of Computer Science, University of Glasgow, 1979) <http://home.mit.bme.hu/~meszaros/edu/onallo/it/valazsik.97/ir/Preface.pdf>, (accessed April 23, 2008).

34. "Current Projects," The Information Retrieval Group, University of Glasgow, (August 30, 2007) <http://ir.dcs.gla.ac.uk/projects.html> (accessed April 23, 2008).

35. "W. Bruce Croft," Department of Computer Science, University of Massachusetts Amherst, <http://ciir.cs.umass.edu/personnel/croft.html> (accessed April 23, 2008).

### Image citations

Cyril Cleverdon: "Awards", Special Interest Group on Information Retrieval, Gerard Salton Award, <http://www.sigir.org/awards/awards.html#salton> (accessed April 23, 2008).

Gerald Salton: "In Memory of: Gerard Salton," Computer Science Department, Cornell University, <http://www.cs.cornell.edu/Info/Department/Annual95/Faculty/Salton.html> (accessed April 23, 2008).

Karen Spärck Jones: "Karen Spärck Jones (1935-2007)," University of Cambridge, Computer Laboratory, Obituaries, <http://www.cl.cam.ac.uk/misc/obituaries/sparck-jones/> (accessed March 26, 2008).

Stephen E. Robertson: "Stephen E. Robertson: Non-Stipendiary Fellow," Girton College, University of Cambridge, <http://girton.etianen.com/fellows-and-staff/non-stipendiary-fellows/s-robertson/> (accessed April 23, 2008).

CJ Rijsbergen young: "C.J. 'Keith' van Rijsbergen," Department of Computer Science, University of Glasgow, <http://www.dcs.gla.ac.uk/~iain/keith/data/images/keith.gif> (accessed April 23, 2008).

CJ Rijsbergen now: "Awards", Special Interest Group on Information Retrieval, Gerard Salton Award, <http://www.sigir.org/awards/awards.html#salton> (accessed April 23, 2008).

Bruce Croft: "W. Bruce Croft," Department of Computer Science, University of Massachusetts Amherst, <http://ciir.cs.umass.edu/personnel/croft.html> (accessed April 23, 2008).