Duality in information retrieval and the hypergeometric distribution

Duality in information retrieval and the hypergeometric distribution

LEO EGGHE * *and* RONALD ROUSSEAU **

* *LUC, Universitaire Campus, B-3590 Diepenbeek, Belgium*
and UIA, Universiteitsplein, 1, 2610, Wilrijk, Belgium
legghe@luc.ac.be

** *KHBO, Zeedijk 101, B-8400 Oostende, Belgium*
and UIA, Universiteitsplein, 1, 2610, Wilrijk, Belgium
Ronald.Rousseau@kh.khbo.be

Abstract

Duality is an important topic in informetrics, especially in connection with the classical informetric laws. Yet, this concept is less studied in information retrieval. It deals with the unification or symmetry between queries and documents, retrieval versus indexing, and relevant versus retrieved documents. These ideas are elaborated in this note and the connection with the hypergeometric distribution is highlighted.

1. INTRODUCTION: DUALITY

Classical informetrics deals with two types of objects and their relation: sources and items, where the former 'produce' the latter. Examples abound and are well known: authors produce articles (Lotka terminology), journals publish articles on a specific subject (Bradford terminology). In linguistics, words (types) 'produce' occurrences in a text (tokens). Here Zipf's or Mandelbrot's "law" is applicable. In scientometrics we consider articles as sources and citations (mentions in other authors' reference list) as items. Note, however, that one can consider articles as sources and the publications presented in these articles' own reference lists as (produced) items. For a detailed description of these informetric regularities we refer to [1]; a more elementary review can be found in [2].

The last mentioned examples exhibit already a duality aspect, namely between citing and cited. But, what does duality mean, here and in general? One well known duality situation is that studied in projective geometry, between lines and points. There it can be shown that any true statement involving straight lines and points can be transformed into another one (not trivially implied by the former one), by interchanging the terms 'points' and 'lines'. It is clear that in this example, duality has a lot to do with a kind of symmetry and not with a form of 'identification' or 'unification': points and lines are different objects. Similarly, in physics, Maxwell's electromagnetic theory shows a remarkable duality: if in Maxwell's equations the electric field E and the magnetic field B are interchanged, and also the electric charge e, and the magnetic charge g, then the equations stay the same.

Formally, duality is defined in the framework of category theory. In this highly abstract mathematical framework, every class of objects and morphisms (a so-called category) has a well-defined dual (or opposite) category. By the duality principle, when a theorem is true, its dual theorem (formulated in the dual category) is automatically also true. Basic references for category theory are [3] and [4]. In [5] we have shown how the theory of Information Production Processes, i.e. the abstract theory of sources, items and their relations [6], can be described in a categorical framework. Here, we will not explicitly use this highly abstract construction, but we will try to stay as close as possible to it, without employing the corresponding mathematical formalism.

In the sequel we will investigate the notion of duality in the context of information retrieval (IR). Here we will focus on the possible duality between documents and queries. Indeed, as the aim of an information retrieval system is to find information items relevant to an information need and as relevance is a kind of similarity relation between the concepts represented by the information item and

those represented by the formulation of the information need (the query), it is not astonishing to discover that the class of possible queries can often be considered the same as or similar to the class of possible representations of information items. Moreover, queries can often be considered as virtual items [7]. We will discuss unification and symmetry, relevant versus retrieved documents, retrieval versus indexing. We will show that the hypergeometric distribution is the "natural" underlying distribution in IR, no matter whether we consider a topic from a retrieval or an indexing point of view. This note is only meant to be a contribution to present ideas for what should be an interesting topic in theoretical information retrieval. As such our discussion is at times heuristic and the same can be said of some definitions.

2. DUALITY IN INFORMATION RETRIEVAL

2.1 Unification or symmetry

Robertson [8] explicitly discusses the concept of duality in IR, considering unification as well as symmetry between queries and documents. Unification occurs e.g. in the vector space model [9]: both queries and documents are modelled as N-vectors, where N is the total number of available index terms. The components of these N-vectors can be real numbers, or real numbers between zero and one (weighted model) or simply the numbers zero or one (discrete model). In the former two cases, the i-th component describes the importance of term i in the document or the query. All possible queries can be visualized by all possible documents, and vice-versa. Hence, in this approach, there is no conceptual difference between queries and documents. Of course, once we have fixed the document set and the set of queries, we can ask what happens if we interchange the terms "document" and "query".

As Robertson argues, one can very well diversify between queries and documents, but in those cases where the IR process consists of matching documents to queries, we can as well match queries to documents. Again the vector space model provides an example. In this model a database consists of n documents, while the query space (the set of all possible queries) is $[0,1]^N$, i.e. any vector $(x_1,...,x_N)$ with $x_i \in [0,1]$ $(i = 1,...,N)$ can be considered as a possible query. Yet, matching a document and a query can be ruled by a symmetric function

$$\text{sim}(D,Q) = \text{sim}(Q,D) \qquad (1)$$

where D denotes any document and Q any query; sim denotes a similarity function such as Salton's cosine measure [10], which is defined in such a way that (1) becomes meaningful. We will comment on this further on (cf equations (2),(3) and (4)).

2.2 General treatment of duality in information retrieval

2.2.1 Duality between queries and documents

Any database will be considered as a document space, denoted as DS. In all practical cases the set DS is finite, but in our model even infinite document spaces are allowed. The set DS contains all document representations under consideration. Similarly, we will denote by QS, the set of all possible query representations. Note the adjective 'possible'. Indeed, since queries are not physical objects, they come into existence as soon as they have been formulated in order to perform an IR action. The set QS can be finite or infinite. There is even more reason here to assume that QS is infinite. So, in general DS $\neq$ QS (as advocated by Bollmann-Sdorra and Raghavan [11]), although equality is certainly not excluded.

DS is obtained as a surjective image of a full-text literature set, through an indexing process. QS, on the other hand, is constructed as a surjective image of an 'imaginary', complex problem database. Its elements come into being every time a user has a (scientific) problem. In Belkin's terminology [12] we could say that this person has a recognized anomalous state of knowledge.

An IR-system is merely a browsing system if there is no way to match documents and queries. This matching is performed through a "similarity" function (take this term in a broad sense), denoted as sim. Symbolically:

$$\text{sim} : DS \times QS \longrightarrow \mathbb{R}^+ : (D,Q) \longrightarrow \text{sim}(D,Q) \qquad (2)$$

where $\mathbb{R}^+$ denotes the positive real numbers (including zero). In plain terms, equation (2) states that the similarity between a document and a query (in that order) is obtained by using the function sim. Now, we want to know what the similarity is between a query and a document (in this order). This can easily be solved by introducing a function sim^*, where sim^* is defined as follows:

$$\text{sim}^* : QS \times DS \longrightarrow \mathbb{R}^+ : (Q,D) \longrightarrow \text{sim}^*(Q,D) \qquad (3)$$

$$\text{with } \text{sim}^*(Q,D) = \text{sim}(D,Q) \qquad (4)$$

for all D $\in$ DS and Q $\in$ QS. Often, in a heuristic way, one deletes the symbol *, and uses only the sim-notation.

This brings us to a first aspect of duality: no matter how different documents and queries are, the similarity between a document D and a query Q is the same as the similarity between this query Q and this document D. This first aspect of duality implies, by (3) and (4), that we can interchange the roles of DS and QS. In this sense, QS can be considered as a document space and DS as a query space. This is analogous to the symmetry aspect as described above and in [8].

2.2.2 Duality between indexing and retrieval

As a consequence of the above we also have a duality between indexing and retrieval. Indexing can be seen as the attachment of a vocabulary to documents according to certain rules. Here the term 'vocabulary' is used in a generic sense: it may refer to keywords, or to any meaningful word in an abstract, or to any other result of attaching a set of words to a full-text document. As an extreme case in can even be the full-text document itself. Hence, it is the physical act of representing full-text documents by elements of DS. Retrieval, on the other hand, can be seen as the attachment of keywords to a problem state (an ASK), leading to a "query-element" in QS. Hence, the duality discussed in Section 2.2.1 leads to a duality between indexing and retrieval.

Some specific remarks can be made here. In indexing a distinction is made between the use of single terms and the use of terms in context. Indexing using combined terms or terms in context is known as precoordination [10]. These combined terms and phrases yield an alternative description of the document. Precoordination is done doing the indexing phase and these terms and phrases can be used as such during the retrieval process. Using single terms, such as keywords or descriptors, leads to postcoordination: the user has to combine different terms when searching for a compound notion, a process which can be done using e.g. proximity operators. Hence, postcoordination is a retrieval action. Because of the above mentioned duality between indexing and retrieval, pre- and post-coordination can be considered as dual features.

Boolean searching is characterized by combining single keywords and forming AND, OR or NOT-queries (or combinations thereof). Such queries are executed in the document space by applying the corresponding intersection ($\cap$), union ($\cup$) or difference (\) operator on subsets of the

document space. So, the equivalence between Boolean set theory and Boolean logic also has its origins in duality.

2.2.3 Duality in IR and mathematical duality

Formulae (2) and (3) yield also a mathematical interpretation in IR. The similarity function sim can be restricted to DS as follows: fix $Q \in QS$ and denote by $sim(\cdot, Q)$ the function

$$sim(\cdot, Q) : DS \longrightarrow \mathbb{R} : D \longrightarrow sim(D,Q) \qquad (5)$$

Often, DS is a vector space (e.g. when components of the vector can take any real value) and the similarity function is a linear function defined on this vector space This is e.g. the case for the inner product used as a similarity measure in the vector space model [9]. In mathematics, this is expressed by saying that sim $(\cdot, Q)$ belongs to the dual vector space of DS, denoted as DS^*. So $sim(\cdot, Q) \in DS^*$. Now, it is clear that one can identify sim $(\cdot, Q)$ and Q (every Q leads to such a linear mapping, while such a mapping leads to a unique query, if the set $\{sim(D, \cdot), D \in DS\}$ separates the points of QS), and hence QS becomes a subspace of DS^*. Mathematically, queries are elements of the dual space of the document space. Note that also the dual statement is true: for every $D \in DS$

$$sim(D,\cdot) : QS \longrightarrow \mathbb{R}^+ : Q \longrightarrow sim(D,Q) \qquad (6)$$

In this way D becomes an element of the dual space of QS.

Not only the weighted vector space model can be considered in this way. If one attaches only the values zero and one to queries and documents, then DS and QS can be identified with $\mathbf{Z}_2^N$ (N is the number of used keywords). Here $\mathbf{Z}_2$ denotes the two-element vector space {0,1} where adding and multiplication are performed as follows: 0+0 = 0, 0+1=1+0=1, 1+1 = 0; 0.0 = 1.0 = 0.1 = 0, 1.1 = 1. Sometimes this way of adding is referred to as counting modulo 2. As clearly $\mathbf{Z}_2^N = (\mathbf{Z}_2^N)^*$ (see appendix for a proof), unweighted retrieval (i.e. a keyword is used (1), or it is not (0)) leads to a special kind of duality between queries and documents, namely unification in a strong mathematical sense.

2.2.4 Duality between retrieved documents and relevant documents

In practical IR there is a serious conceptual difference between retrieved and relevant documents. Indeed, their relation forms the basis of the classical recall-precision performance measures. Of course, in a perfect world the retrieved documents should be identical with the relevant ones (e.g. with

respect to a query Q). If one could work with the full text documents and the 'full' problem description, this could be achieved by browsing through the complete database. However, as explained in subsection 2.2.1, documents in DS, as well as queries in QS are surjective images of these full text originals and therefore one cannot expect the set of retrieved documents to be the same as the set of relevant documents.

Moreover, after retrieval, the retrieved documents are known to the investigator while not all relevant ones are known. Yet, there is a duality between these two sets. Retrieved documents are the result of a retrieval action, relevant documents are the result of an indexing action. We admit this needs some explanation. Suppose then that we have no IR software available and, hence, that, to find documents in a database, which then looks more like an (electronic) filing cabinet, we have to check the entire database. Based on our need for knowledge, we are able to find all the relevant documents (here we assume that indexing is done in a detailed manner so that it is clear what the corresponding full-text document is all about). On the other hand it needs no explanation that retrieved documents are the result of a retrieval action. In practice, the limitations of such an action result in a set that is possibly different from the set of relevant documents.

Based on the duality between indexing and retrieval (Subsection 2.2.2) we have hence reached a dual explanation of relevance and retrievals.

One can, finally, also remark that a query, composed by the researcher as being relevant to a certain scientific problem, results in a set of retrieved documents So a relevant query yields (not necessarily relevant) retrieved documents.

2.2.5 Duality and the hypergeometric distribution

It is only recently that the hypergeometric distribution was highlighted in connection with IR. Besides an 'older' article by Wilbur [13], which uses the hypergeometric distribution as a tool to describe the probability that a document is related to a query, we only know of the recent articles by Shaw et al [14] and Egghe & Rousseau [15]. As we feel that this distribution is basic to all quantitative approaches to IR, we present the simple argument which leads to its use.

Fig.1 represents an abstract picture of information retrieval: there is a database DS in which we show a set A obtained as the result of a retrieval action, based on a query $Q \in QS$. So we put $A = ret_Q$.

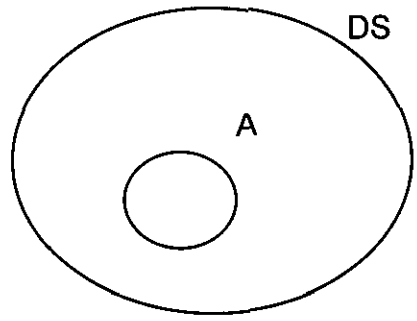

Fig.1 Retrieval of a set A = ret_Q from a database Ω = DS.

Let us denote the number of elements in DS by n, and the number of elements in A by m. Suppose that the query Q implies the existence of k relevant documents in DS. A basic problem of IR is: what is the probability of retrieving p of these k relevant documents ($0 \leq p \leq k$). In other words: what is the chance to find, in set A, exactly p of the k relevant documents?

One argument is very simple: the probability is the quotient of the total number of favourable cases and the total number of possible cases. Here, favourable means that set A contains p relevant documents out of the k in the database available relevant ones. These values are calculated as follows: since A must contain p relevant documents, there are $\binom{m}{p}$ possible cases, multiplied by $\binom{n-m}{k-p}$ possible cases of k-p relevant documents outside A. Hence the numerator is

$$\binom{m}{p}\binom{n-m}{k-p} \qquad (7)$$

The total number of possible cases to have k relevant documents among n is $\binom{n}{k}$. Hence, the probability sought is

$$P(n,m,k,p) = \frac{\binom{m}{p}\binom{n-m}{k-p}}{\binom{n}{k}} \qquad (8)$$

It can readily be checked that (8) describes a discrete distribution, since

$$\sum_{p=0}^{k} \binom{m}{p}\binom{n-m}{k-p} = \binom{n}{k} \qquad (9)$$

Formula (9) appears e.g. in [16] and is moreover easy to prove by induction. The distribution (8) is the classical hypergeometric distribution [17].

We stress the importance of the hypergeometric distribution for IR: it calculates the basic uncertainties that appear in retrieval actions, namely the probabilities to find relevant documents.

Independent of Egghe and Rousseau [14], and prior to them, Shaw et al. [15] produced the following 'dual' argument. Using the notation of this article (and of [14]), there are $\binom{k}{p}$ possible ways of picking p relevant documents out of k, and $\binom{n-k}{m-p}$ ways of picking m-p non-relevant documents out of n-k. Hence, the probability to find p relevant documents out of k is:

$$P^*(n,m,k,p) = \frac{\binom{k}{p}\binom{n-k}{m-p}}{\binom{n}{m}} \qquad (10)$$

Indeed, in total, there are $\binom{n}{m}$ possible selections of a set with m documents out of DS, having n documents. At first sight, (10) is completely different from (8), and it seems that it is not even a hypergeometric distribution, since the variable p is an element of {0,...,k} and not of {0,...,n}. But a straightforward calculation, using only the definition of binomial coefficients, shows that

$$P(n,m,k,p) = P^*(n,m,k,p)$$

Now (10) follows from (8) (and vice-versa) by interchanging the symbols k and m, i.e. by replacing the set A of retrieved documents by the set of relevant documents.

This shows that the hypergeometric distribution *is independent of the relevant-retrieved duality.*

SUMMARY

We have highlighted different duality aspects of IR systems such as query-document, retrieval-indexing and relevant-retrieved. Also pre and post coordination have been put in a dual relation.

Further, Boolean logic has been dually related with Boolean set operations. The mathematical notion of duality has been touched upon by considering dual vector spaces. Finally, it has been noted that the hypergeometric distribution is basic in IR and is independent of the dual switch relevant-retrieved.

REFERENCES

1. EGGHE,L. *and* ROUSSEAU,R. *Introduction to informetrics*. Amsterdam: Elsevier, 1990.
2. ROUSSEAU,R. *and* ROUSSEAU,S. Informetric distributions: a tutorial review. *The Canadian Journal of Information and Library Science/Revue canadienne des sciences de l'information et de bibliothéconomie, 18,* 1993, 51-63.
3. MAC LANE,S. *Categories for the working mathematician*. New York: Springer-Verlag, 1971.
4. HERRLICH,H. *and* STRECKER,G.E. *Category theory*. Boston: Allyn & Bacon, 1973.
5. ROUSSEAU,R. Category theory and informetrics: information production processes. *Scientometrics, 25,* 1992, 77-87.
6. EGGHE,L. The duality of informetric systems with applications to the empirical laws. *Journal of Information Science, 16,* 1990, 17-27.
7. BARTCHI,M. An overview of information retrieval subjects. *Computer, 18,* May 1985, 67-84.
8. ROBERTSON,S. Query-document symmetry and dual models. *Journal of Documentation, 50,* 1994, 233-238.
9. RAGHAVAN,V.V. *and* WONG,S.K.M. A critical analysis of vector space model for information retrieval. *Journal of the American Society for Information Science, 37,* 1986, 279-287.
10. SALTON,G. *and* McGILL,M.J. *Introduction to modern information retrieval.* New York:McGraw-Hill, 1983.
11. BOLLMANN-SDORRA,P. and RAGHAVAN, V.V. On the delusiveness of adopting a common space for modeling IR objects: are queries documents? *Journal of the American Society for Information Science, 44,* 1993, 579-587.
12 BELKIN,N.J. Anomalous states of knowledge as a basis for information retrieval. *Canadian Journal of Information Science, 5,* 1980, 133-143.
13. WILBUR, W.J. Retrieval testing with hypergeometric document models. *Journal of the American Society for Information Science, 44,* 1993, 340-351.

14. EGGHE,L. *and* ROUSSEAU,R. A theoretical study of recall and precision using a topological approach to information retrieval. Preprint, 1997.

15. SHAW,W.M.Jr, BURGIN,R. *and* HOWELL,P. Performance standards and evaluations in IR test collections: vector-space and other retrieval models. *Information Processing and Management, 33,* 1997, 15-36.

16. GRADSHTEYN,I.S., RYZHIK,I.M. *and* JEFFREY,A. *Table of integrals, series and products.* New York: Academic Press, 1980.

17. FELLER,W. *An introduction to probability theory and its applications.* New York:Wiley, 1968.

Appendix: Proof of : $\mathbf{Z}_2^N = (\mathbf{Z}_2^N)^*$

The dual space of $\mathbf{Z}_2$ (considered as a vector space over the field $\mathbf{Z}_2$) consists of all linear mapping from $\mathbf{Z}_2$ to $\mathbf{Z}_2$. However, a linear mapping always sends zero to zero, so there are only two alternatives: or 1 is send to 0, or 1 is send to 1. In the first case we have the zero map (0), in the second the identity map. This proves that $\mathbf{Z}_2$ is self-dual, i.e. $\mathbf{Z}_2 = \mathbf{Z}_2^*$.

Next, we will show that $(\mathbf{Z}_2^*)^N = (\mathbf{Z}_2^N)^*$

Let i_j be the canonical injection of $\mathbf{Z}_2$ onto the j-th component of $\mathbf{Z}_2^N$: $a \longrightarrow (0,0,\ldots,a,0,\ldots 0)$. If now $f \in (\mathbf{Z}_2^N)^*$ is given, define then $(f_j) \in (\mathbf{Z}_2^*)^N$ as: $f_j = f \circ i_j$, $j = 1,\ldots,\mathbf{N}$. Conversely, if $(g_j) \in (\mathbf{Z}_2^*)^N$ is given, define $g \in (\mathbf{Z}_2^N)^*$ by $g(a_1,a_2,\ldots,a_N) = g_1(a_1) + \ldots + g_N(a_N)$. It is clear that the f_j's and g are linear mappings and that $(f_j)_j$ and g introduce a vector isomorphism between $(\mathbf{Z}_2^*)^N$ and $(\mathbf{Z}_2^N)^*$.

Finally, combining the two results shows that $\mathbf{Z}_2^N = (\mathbf{Z}_2^N)^*$.

Aan Prof. Dr. S.E. Robertson
Editor Journal of Documentation
Department of Information Science
City University, London, UK

Diepenbeek, February 27, 1997
BIB/LE/970022

Dear Steve,

Please find enclosed three copies of my paper (co-authored with R. Rousseau) entitled "Duality in information retrieval and the hypergeometric distribution". We propose it to you since you are one of the editors of Journal of Documentation but also because of your interest in duality in IR.

We would like to submit it to the above mentioned journal. Thank you for considering this paper.

Yours sincerely,

Prof. Dr. L. Egghe
LUC
Universitaire Campus
B-3590 Diepenbeek
Belgium

JOURNAL OF DOCUMENTATION

Editor R.T. Kimber Assistant Editor A.M. Adams
Queen's University Science Library Chlorine Gardens Belfast
Northern Ireland BT9 5EQ Tel.: (01232) 335442 Fax: (01232) 382636 Email: j.doc@qub.ac.uk

14 March 1997

Professor L Egghe
Limburgs Universitair Centrum
Universitaire Campus
B-3590 Diepenbeek
Belgium

Dear Professor Egghe,

Duality in information retrieval and the hypergeometric distribution

Stephen Robertson gave me this paper yesterday - many many thanks. It's nice to hear from you again.

The paper will be refereed in our usual way. I should hope to have reports in time for discussion at the next Editorial Board meeting which is in mid June and I shall be in touch with you again after that.

With best wishes.

Yours sincerely,

Amber M Adams

Published by Aslib, The Association for Information Management 20-24 Old Street London EC1V 9AP
Tel.: (0171) 2534488 Fax: (0171) 4300514 Email: publications@aslib.co.uk WWW: http://www.aslib.co.uk/aslib/

Duality in information retrieval and the hypergeometric distribution

LEO EGGHE * *and* RONALD ROUSSEAU **

* *LUC, Universitaire Campus, B-3590 Diepenbeek, Belgium*
and UIA, Universiteitsplein, 1, 2610, Wilrijk, Belgium
legghe@luc.ac.be

** *KHBO, Zeedijk 101, B-8400 Oostende, Belgium*
and UIA, Universiteitsplein, 1, 2610, Wilrijk, Belgium
Ronald.Rousseau@kh.khbo.be

Abstract

Duality is an important topic in informetrics, especially in connection with the classical informetric laws. Yet, this concept is less studied in information retrieval. It deals with the unification or symmetry between queries and documents, search formulation versus indexing, and relevant versus retrieved documents. These ideas are elaborated in this note and the connection with the hypergeometric distribution is highlighted.

1. INTRODUCTION: DUALITY

Classical informetrics deals with two types of objects and their relation: sources and items, where the former 'produce' the latter. Examples abound and are well known: authors produce articles (Lotka terminology), journals publish articles on a specific subject (Bradford terminology). In linguistics, words (types) 'produce' occurrences in a text (tokens). Here Zipf's or Mandelbrot's "law" is applicable. In scientometrics we consider articles as sources and citations (mentions in other authors' reference list) as items. Note, however, that one can consider articles as sources and the publications presented in these articles' own reference lists as (produced) items. For a detailed description of these informetric regularities we refer to [1]; a more elementary review can be found in [2].

The last mentioned examples exhibit already a duality aspect, namely between citing and cited. But, what does duality mean, here and in general? One well known duality situation is that studied in projective geometry, between lines and points. There it can be shown that any true statement involving straight lines and points can be transformed into another one (not trivially implied by the former one), by interchanging the terms 'points' and 'lines'. It is clear that in this example, duality has a lot to do with a kind of symmetry and not with a form of 'identification' or 'unification': points and lines are different objects. Similarly, in physics, the vacuum Maxwell equations for the electric and magnetic fields **E** and **B**, show a remarkable duality: substituting **B** for **E** and **–E** for **B** leaves the equations invariant.

Formally, duality is defined in the framework of category theory. In this highly abstract mathematical framework, every class of objects and morphisms (a so-called category) has a well-defined dual (or opposite) category. By the duality principle, when a theorem is true, its dual theorem (formulated in the dual category) is automatically also true. Basic references for category theory are [3] and [4]. In [5] we have shown how the theory of Information Production Processes, i.e. the abstract theory of sources, items and their relations [6], can be described in a categorical framework. Here, we will not explicitly use this highly abstract construction, but we will try to stay close to it, without employing the corresponding mathematical formalism.

In the sequel we will investigate the notion of duality in the context of information retrieval (IR). Here we will focus on the possible duality between documents and queries. Indeed, as the aim of an information retrieval system is to find information items relevant to an information need and as relevance is a kind of similarity relation between the concepts represented by the information item and those represented by the formulation of the information need (the query), it is not astonishing to

discover that the class of possible queries can often be considered the same as or similar to the class of possible representations of information items. Moreover, queries can often be considered as virtual items [7]. We will discuss unification and symmetry, relevant versus retrieved documents, search formulation versus indexing. We will show that the hypergeometric distribution is the "natural" underlying distribution in IR, no matter whether we consider a topic from a retrieval or an indexing point of view. This note is only meant to be a contribution to present ideas for what should be an interesting topic in theoretical information retrieval. As such our discussion is at times heuristic and the same can be said of some definitions.

2. DUALITY IN INFORMATION RETRIEVAL

2.1 Unification or symmetry

Robertson [8] explicitly discusses the concept of duality in IR, considering unification as well as symmetry between queries and documents. Unification occurs e.g. in the vector space model [9]: both queries and documents are modelled as N-vectors, where N is the total number of available index terms. The components of these N-vectors can be real numbers, or real numbers between zero and one (weighted model) or simply the numbers zero or one (discrete model). In the former two cases, the i-th component describes the importance of term i in the document or the query. All possible queries can be visualised by all possible documents, and vice-versa. Hence, in this approach, there is no conceptual difference between queries and documents. Of course, once we have fixed the document set and the set of queries, we can ask what happens if we interchange the terms "document" and "query".

As Robertson argues, one can very well diversify between queries and documents, but in those cases where the IR process consists of matching documents to queries, we can as well match queries to documents. Again the vector space model provides an example. In this model a database consists of n documents, while the query space (the set of all possible queries) is $[0,1]^N$, i.e. any vector $(x_1,...,x_N)$ with $x_i \in [0,1]$ $(i = 1,...,N)$ can be considered as a possible query. Yet, matching a document and a query can be ruled by a symmetric function

$$sim(D,Q) = sim(Q,D) \qquad (1)$$

where D denotes any document and Q any query; sim denotes a similarity function such as Salton's cosine measure [10], which is defined in such a way that (1) becomes meaningful. We will comment on this further on (cf equations (2),(3) and (4)).

2.2 General treatment of duality in information retrieval

2.2.1 Duality between queries and documents

Any database will be considered as a document space, denoted as DS. In all practical cases the set DS is finite, but in our model even infinite document spaces are allowed. The set DS contains all document representations under consideration. Similarly, we will denote by QS, the set of all possible query representations. Note the adjective 'possible'. Indeed, since queries are not physical objects, they come into existence as soon as they have been formulated in order to perform an IR action. The set QS can be finite or infinite. There is even more reason here to assume that QS is infinite. So, in general DS ≠ QS (as advocated by Bollmann-Sdorra and Raghavan [11]), although equality is certainly not excluded.

DS is obtained as a surjective image of a full-text literature set, through an indexing process. QS, on the other hand, is constructed as a surjective image of an 'imaginary', complex problem database. Its elements come into being every time a user has a (scientific) problem. In Belkin's terminology [12] we could say that this person has a recognised anomalous state of knowledge.

An IR-system is merely a browsing system if there is no way to match documents and queries. This matching is performed through a "similarity" function (take this term in a broad sense), denoted as sim. Symbolically:

$$\text{sim} : DS \times QS \longrightarrow \mathbb{R}^+ : (D,Q) \longrightarrow \text{sim}(D,Q) \qquad (2)$$

where $\mathbb{R}^+$ denotes the positive real numbers (including zero). In plain terms, equation (2) states that the similarity between a document and a query (in that order) is obtained by using the function sim. Now, we want to know what the similarity is between a query and a document (in this order). This can easily be solved by introducing a function sim^*, where sim^* is defined as follows:

$$\text{sim}^* : QS \times DS \longrightarrow \mathbb{R}^+ : (Q,D) \longrightarrow \text{sim}^*(Q,D) \qquad (3)$$

$$\text{with } \text{sim}^*(Q,D) = \text{sim}(D,Q) \qquad (4)$$

for all $D \in DS$ and $Q \in QS$. Often, in a heuristic way, one deletes the symbol *, and uses only the sim-notation.

This brings us to a first aspect of duality: no matter how different documents and queries are, the similarity between a document D and a query Q can be defined to be the same as the similarity between this query Q and this document D. This first aspect of duality implies, by (3) and (4), that we can interchange the roles of DS and QS. In this sense, QS can be considered as a document space and DS as a query space. This is analogous to the symmetry aspect as described above and in [8].

2.2.2 Duality between indexing and search formulation

As a consequence of the above we also have a duality between indexing and search formulation. Indexing can be seen as the attachment of a vocabulary to documents according to certain rules. Here the term 'vocabulary' is used in a generic sense: it may refer to keywords, or to any meaningful word in an abstract, or to any other result of attaching a set of words to a full-text document. As an extreme case in can even be the full-text document itself. Hence, it is the physical act of representing full-text documents by elements of DS. Search formulation, on the other hand, can be seen as the attachment of keywords to a problem state (an ASK), leading to a "query-element" in QS. Hence, the duality discussed in Section 2.2.1 leads to a duality between indexing and search formulation.

Some specific remarks can be made here. In indexing a distinction is made between the use of single terms and the use of terms in context. Indexing using combined terms or terms in context is known as precoordination [10]. These combined terms and phrases yield an alternative description of the document. Precoordination is done during the indexing phase and these terms and phrases can be used as such during the search formulation process. Using single terms, such as keywords or descriptors, leads to postcoordination during retrieval: the user has to combine different terms when searching for a compound notion, a process which can be done using e.g. proximity operators. Hence, postcoordination is a search formulation action. Because of the above mentioned duality between indexing and search formulation, pre- and post-coordination can be considered as dual features.

Boolean search is characterised by combining single keywords and forming AND, OR or NOT-queries (or combinations thereof). Such queries are executed in the document space by applying the corresponding intersection ($\cap$), union ($\cup$) or difference ($\setminus$) operator on subsets of the document space. So, the equivalence between Boolean set theory and Boolean logic also has its origins in duality: Boolean logic in the query formulation, and Boolean set theory in the selection of the corresponding documents.

2.2.3 Duality in IR and mathematical duality

Formulae (2) and (3) yield also a mathematical interpretation in IR. The similarity function sim can be restricted to DS as follows: fix $Q \in QS$ and denote by $sim(\cdot, Q)$ the function

$$sim(\cdot, Q) : DS \longrightarrow \mathbb{R} : D \longrightarrow sim(D,Q) \qquad (5)$$

Often, DS is a vector space (e.g. when components of the vector can take any real value) and the similarity function is a linear function defined on this vector space This is, e.g., the case for the inner product used as a similarity measure in the vector space model [9], [10]. In mathematics, the linear case is expressed by saying that sim $(\cdot, Q)$ belongs to the dual vector space of DS, denoted as DS^*. So sim $(\cdot, Q) \in DS^*$. Now, it is clear that one can identify sim $(\cdot, Q)$ and Q (every Q leads to such a linear mapping, while such a mapping leads to a unique query, at least if the set $\{sim(D, \cdot), D \in DS\}$ separates the points of QS). This expression means that if $sim(D, Q_1) = sim(D,Q_2)$, for all $D \in DS$, then $Q_1 = Q_2$. By this construction QS becomes a subspace of DS^*. Mathematically, queries are elements of the dual space of the document space. Note that also the dual statement is true: for every $D \in DS$ there exists a mapping

$$sim(D,\cdot) : QS \longrightarrow \mathbb{R}^+ : Q \longrightarrow sim(D,Q) \qquad (6)$$

In this way D becomes an element of the dual space of QS.

Not only the weighted vector space model can be considered in this way. If one attaches only the values zero and one to queries and documents, then DS and QS can be identified with $\mathbf{Z}_2^N$ (N is the number of used keywords). Here $\mathbf{Z}_2$ denotes the two-element vector space {0,1} where adding and multiplication are performed as follows: 0+0 = 0, 0+1=1+0=1, 1+1 = 0; 0.0 = 1.0 = 0.1 = 0, 1.1 = 1. Sometimes this way of adding is referred to as counting modulo 2. As clearly $\mathbf{Z}_2^N = (\mathbf{Z}_2^N)^*$ (see appendix for a proof), unweighted retrieval (i.e. a keyword is used (1), or it is not (0)) leads to a special kind of duality between queries and documents, namely unification in a strong mathematical sense.

2.2.4 Duality between retrieved documents and relevant documents

In practical IR there is a serious conceptual difference between retrieved and relevant documents. Indeed, their relation forms the basis of the classical recall-precision performance measures. Of course, in a perfect world the retrieved documents should be identical with the relevant ones (e.g. with respect to a query Q). If one could work with the full text documents and the 'full' problem description,

this could be achieved by browsing through the complete database. However, as explained in subsection 2.2.1, documents in DS, as well as queries in QS are surjective images of these full text originals and therefore one cannot expect the set of retrieved documents to be the same as the set of relevant documents.

Moreover, after retrieval, the retrieved documents are known to the investigator while not all relevant ones are known. Yet, there is a duality between these two sets. Retrieved documents are the result of a search formulation, relevant documents are the result of an indexing action. We admit this needs some explanation. Suppose then that we have no IR software available and, hence, that, to find documents in a database, which then looks more like an (electronic) filing cabinet, we have to check the entire database. Based on our need for knowledge, we are able to find all the relevant documents (here we assume that indexing is done in a detailed manner so that it is clear what the corresponding full-text document is all about). On the other hand it needs no explanation that retrieved documents are the result of a search formulation. In practice, the limitations of such an action result in a set that is possibly different from the set of relevant documents.

Based on the duality between indexing and search formulation (Subsection 2.2.2) we have hence reached a dual explanation of relevance and retrievals.

One can, finally, also remark that a query, composed by the researcher as being relevant to a certain scientific problem, results in a set of retrieved documents So a relevant query yields (not necessarily relevant) retrieved documents.

2.2.5 Duality and the hypergeometric distribution

It is only recently that the hypergeometric distribution was highlighted in connection with IR. Besides an 'older' article by Wilbur [13], which uses the hypergeometric distribution as a tool to describe the probability that a document is related to a query, we only know of the recent articles by Shaw et al [14] and Egghe & Rousseau [15]. As we feel that this distribution is basic to all quantitative approaches to IR, we present the simple argument which leads to its use.

Fig.1 represents an abstract picture of information retrieval: there is a database DS in which we show a set A obtained as the result of a retrieval action, based on a query $Q \in QS$. So we put $A = ret_Q$.

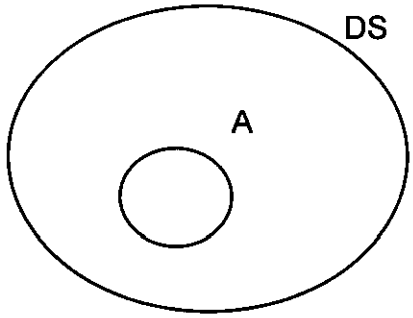

Fig.1 Retrieval of a set A = ret_Q from a database Ω = DS.

Let us denote the number of elements in DS by n, and the number of elements in A by m. Suppose that the query Q implies the existence of k relevant documents in DS. A basic problem of IR is: what is the probability of retrieving p of these k relevant documents ($0 \le p \le k$). In other words: what is the chance to find, in set A, exactly p of the k relevant documents?

One argument is very simple: the probability is the quotient of the total number of favourable cases and the total number of possible cases. Here, favourable means that set A contains p relevant documents out of the k in the database available relevant ones. These values are calculated as follows: since A must contain p relevant documents, there are $\binom{m}{p}$ possible cases, multiplied by $\binom{n-m}{k-p}$ possible cases of k-p relevant documents outside A. Hence the numerator is

$$\binom{m}{p}\binom{n-m}{k-p} \qquad (7)$$

The total number of possible cases to have k relevant documents among n is $\binom{n}{k}$. Hence, the probability sought is

$$P(n,m,k,p) = \frac{\binom{m}{p}\binom{n-m}{k-p}}{\binom{n}{k}} \qquad (8)$$

It can readily be checked that (8) describes a discrete distribution, since

$$\sum_{p=0}^{k} \binom{m}{p}\binom{n-m}{k-p} = \binom{n}{k} \qquad (9)$$

Formula (9) appears e.g. in [16] and is moreover easy to prove by induction. The distribution (8) is the classical hypergeometric distribution [17].

We stress the importance of the hypergeometric distribution for IR: it calculates the *basic uncertainties* that appear in retrieval actions, namely the probabilities to find relevant documents.

Independent of Egghe and Rousseau [14], and prior to them, Shaw et al. [15] produced the following 'dual' argument. Using the notation of this article (and of [14]), there are $\binom{k}{p}$ possible ways of picking p relevant documents out of k, and $\binom{n-k}{m-p}$ ways of picking m-p non-relevant documents out of n-k. Hence, the probability to find p relevant documents out of k is:

$$P^*(n,m,k,p) = \frac{\binom{k}{p}\binom{n-k}{m-p}}{\binom{n}{m}} \qquad (10)$$

Indeed, in total, there are $\binom{n}{m}$ possible selections of a set with m documents out of DS, having n documents. At first sight, (10) is completely different from (8), and it seems that it is not even a hypergeometric distribution, since the variable p is an element of {0,...,k} and not of {0,...,n}. But a straightforward calculation, using only the definition of binomial coefficients, shows that

$$P(n,m,k,p) = P^*(n,m,k,p)$$

Now (10) follows from (8) (and vice-versa) by interchanging the symbols k and m, i.e. by replacing the set A of retrieved documents by the set of relevant documents. This shows that the hypergeometric distribution *is independent of the relevant-retrieved duality.*

We finally mention that the effect of pure random searching is studied in [14], which also includes graphs of the results.

SUMMARY

We have highlighted different duality aspects of IR systems such as query-document, search formulation-indexing and relevant-retrieved. Further, Boolean logic has been dually related with Boolean set operations. The mathematical notion of duality has been touched upon by considering dual vector spaces. Finally, it has been noted that the hypergeometric distribution is basic in IR. Moreover, the occurrence of this distribution is independent of the duality between the notions relevant and retrieved.

REFERENCES

1. Egghe, L. *and* Rousseau, R. *Introduction to informetrics.* Amsterdam: Elsevier, 1990.
2. Rousseau, R. *and* Rousseau, S. Informetric distributions: a tutorial review. *The Canadian Journal of Information and Library Science/Revue canadienne des sciences de l'information et de bibliothéconomie, 18,* 1993, 51-63.
3. Mac Lane, S. *Categories for the working mathematician.* New York: Springer-Verlag, 1971.
4. Herrlich, H. *and* Strecker, G.E. *Category theory.* Boston: Allyn & Bacon, 1973.
5. Rousseau, R. Category theory and informetrics: information production processes. *Scientometrics, 25,* 1992, 77-87.
6. Egghe, L. The duality of informetric systems with applications to the empirical laws. *Journal of Information Science, 16,* 1990, 17-27.
7. Bärtchi, M. An overview of information retrieval subjects. *Computer, 18,* May 1985, 67-84.
8. Robertson, S. Query-document symmetry and dual models. *Journal of Documentation, 50,* 1994, 233-238.
9. Raghavan, V.V. *and* Wong, S.K.M. A critical analysis of vector space model for information retrieval. *Journal of the American Society for Information Science, 37,* 1986, 279-287.
10. Salton, G. *and* McGill, M.J. *Introduction to modern information retrieval.* New York:McGraw-Hill, 1983.
11. Bollmann-Sdorra, P. and Raghavan, V.V. On the delusiveness of adopting a common space for modeling IR objects: are queries documents? *Journal of the American Society for Information Science, 44,* 1993, 579-587.

12 Belkin, N.J. Anomalous states of knowledge as a basis for information retrieval. *Canadian Journal of Information Science, 5,* 1980, 133-143.

13. Wilbur, W.J. Retrieval testing with hypergeometric document models. *Journal of the American Society for Information Science, 44,* 1993, 340-351.

14. Egghe, L. *and* Rousseau, R. A theoretical study of recall and precision using a topological approach to information retrieval. LUC-preprint, 1997 (submitted for publication).

15. Shaw, W.M.Jr, Burgin, R. *and* Howell, P. Performance standards and evaluations in IR test collections: vector-space and other retrieval models. *Information Processing and Management, 33,* 1997, 15-36.

16. Gradshteyn, I.S., Ryzhik, I.M. *and* Jeffrey, A. *Table of integrals, series and products.* New York: Academic Press, 1980.

17. Feller, W. *An introduction to probability theory and its applications.* New York:Wiley, 1968.

Appendix: Proof of : $\mathbf{Z}_2^N = (\mathbf{Z}_2^N)^*$

The dual space of $\mathbf{Z}_2$ (considered as a vector space over the field $\mathbf{Z}_2$) consists of all linear mapping from $\mathbf{Z}_2$ to $\mathbf{Z}_2$. However, a linear mapping always sends zero to zero, so there are only two alternatives: or 1 is send to 0, or 1 is send to 1. In the first case we have the zero map (0), in the second the identity map. This proves that $\mathbf{Z}_2$ is self-dual, i.e. $\mathbf{Z}_2 = \mathbf{Z}_2^*$.

Next, we will show that $(\mathbf{Z}_2^*)^N = (\mathbf{Z}_2^N)^*$

Let i_j be the canonical injection of $\mathbf{Z}_2$ onto the j-th component of $\mathbf{Z}_2^N$: $a \longrightarrow (0,0,\ldots,a,0,\ldots 0)$. If now $f \in (\mathbf{Z}_2^N)^*$ is given, define then $(f_j) \in (\mathbf{Z}_2^*)^N$ as: $f_j = f \circ i_j$, $j = 1,\ldots,N$. Conversely, if $(g_j) \in (\mathbf{Z}_2^*)^N$ is given, define $g \in (\mathbf{Z}_2^N)^*$ by $g(a_1,a_2,\ldots,a_N) = g_1(a_1) + \ldots + g_N(a_N)$. It is clear that the f_j's and g are linear mappings and that $(f_j)_j$ and g introduce a vector isomorphism between $(\mathbf{Z}_2^*)^N$ and $(\mathbf{Z}_2^N)^*$.

Finally, combining the two results shows that $\mathbf{Z}_2^N = (\mathbf{Z}_2^N)^*$.