Information retrieval theory and design based on a model of the user's concept relations*

Matthew B. Koll

6.1 Information retrieval systems as models of human behaviour

6.1.1 Introduction

Viewing information retrieval (IR) systems as models of human assessment of the similarity between requests and documents (SRD)[†] contributes to development of theory for IR and can aid in development of IR systems. This chapter reports the development and testing of a theory of IR based on this system-as-model (SAM) view[‡]. Implications for IR research and development are then considered.

The SAM theory of IR is an expansion of the present model used in IR research. That model, and its expansion into a theory of IR, are described in the two sections below[§].

6.1.2 The traditional model in IR research

The model under which most IR research takes place may be called the system-as-tool (SAT) model. Here the IR system itself is the focus of the research, its components being studied in order to improve the system's usefulness.

The traditional model of IR systems is represented by the Bookstein–Cooper model (Bookstein and Cooper, 1976). *Figure 6.1(a)* is based on that model. Bookstein and Cooper show their model to apply to a wide range of IR systems, from sophisticated automatic systems to ordinary card catalogues.

---

* This work is a contribution of the National Bureau of Standards and is not subject to copyright.
† Owing to confusing definitions and usage of 'relevance', the term SRD (for Similarity between Request and Document) will be used here. SRD corresponds most closely to Swanson's (1977) definition of relevance and the general notion of aboutness (see Maron, 1977). SRD refers to the similarity between the subjects or contents of a document and request.
‡ The development of the SAM theory, as well as most of the work reported here, was conducted as part of a doctoral dissertation (Koll, 1979b).
§. The use of 'model' and 'theory' here is based on Boring's continuum of models-theories (Boring, 1953). The point, on his continuum, where models become theories is where the constructs in the model are hoped to be real, and where the model-theory is thought of as an explanation of the phenomenon, not just a tool for predicting outcomes.
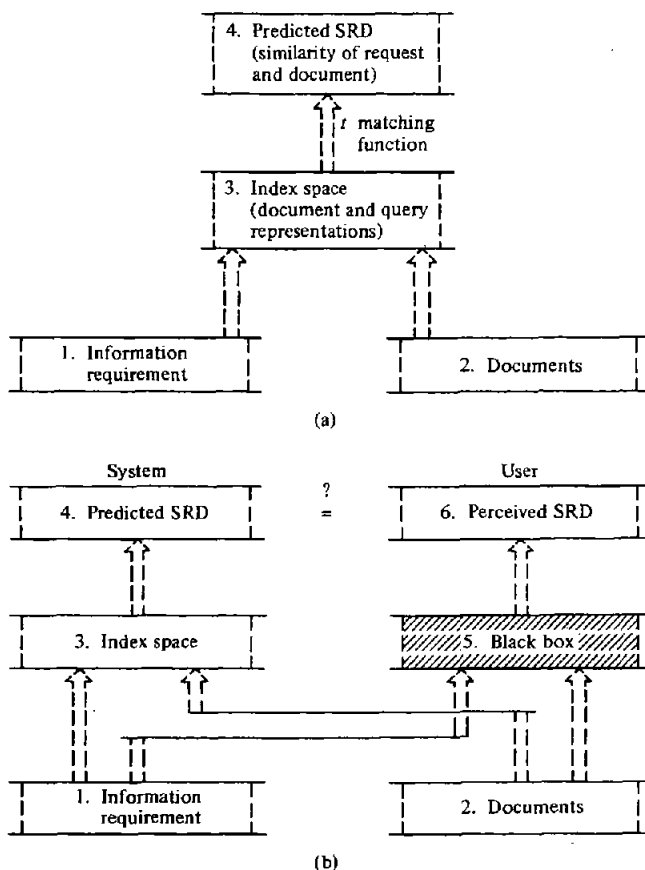
Figure 6.1 (a) General model of IR systems. (b) (SAT) General model of phenomena in IR research (research within this model consists of manipulating components of the left-hand side—i.e. index space— to try to improve the match between boxes 4 and 6)

Figure 6.1(b) is drawn to represent the model under which most IR research is done at present. In this research framework, an IR system is evaluated as a *predictive* model of human judgements (usually relevance). Such research makes no explicit claims about cognitive structures or processes used by the human in making decisions about document relevance, even though the system is trying to predict the outcome of those decisions.

6.1.3   Limitations

The present model underlying IR research may be criticised on three general grounds. (1) It does not provide a strong theoretical framework for research (2) It does not facilitate critical testing of lower-level hypotheses and techniques. (3) It does not provide a basis for advancing IR from symbol-based to concept-based.

(1)   While the system must serve as *at least* a predictive model of human
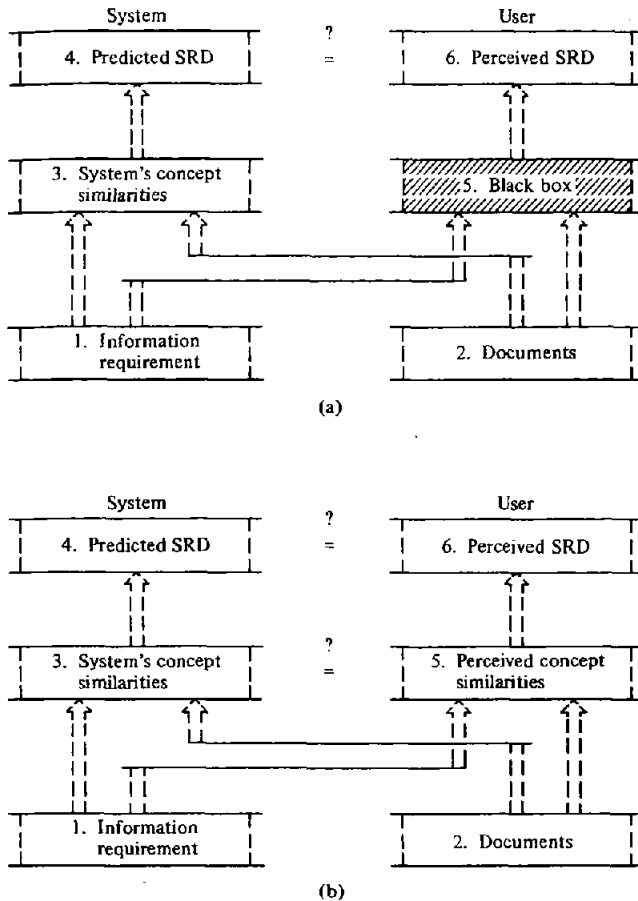
Figure 6.2 (a) (SAT) General model of phenomena in IR research
modified by constraints on syntax, semantics, query representation
and matching function. (b) (SAM) General model of phenomena in
IR research (research within this model may look at factors affecting
the 3 = 5 match as well as the 4 = 6 match. The 4 = 6 match is
dependent on the 3 = 5 match)

behaviour, it could do more. By filling in the black box (*see Figures 2a* and
2b) the IR theorists could attempt to explain, not just predict, the values
placed on documents by users.

The SAT model is weak in that it makes no prohibitions. An important
function of a theory is to tell us what things are *not* going to occur
(Popper, 1965). A theory of IR should define deficient classes of systems,
precluded by theory from outperforming other classes of systems.

(2)   The SAT model only allows comparison of model and subject (that is,
system and user) at the ultimate stage of performance: evaluation of the
documents' relation to a query. When the only dependent measure
(retrieval effectiveness) is so far removed from the independent variables
being manipulated (for example, form of document representation),
uncontrolled factors will often affect the results.

(3)   The SAT model does not make reference to the way people understa text. As pointed out repeatedly by Bar-Hillel (e.g. 1975), Maron a Kuhns (1960) and Robertson (1977), symbol matching is not sufficie Significant progress in IR will not be made until systems 'understai what a document is about. A theory of IR must contain a model of h people come to understand the contents of documents, or at least of h they differentiate documents with respect to a request.

6.1.4   Development of a theory of IR

6.1.4.1   Base assumptions

The development of the SAM theory from the SAT model can be seen modification of *Figure 1(b)* to *Figure 2(a)*. This change consists of constrai on the syntax and semantics of the index space, thus controlling all index sp variables except concept similarity. These constraints are consistent both w state of the art practices in fully automatic document retrieval systems a with traditional assumptions made in IR (*see* Koll, 1979b, page 16).

The shift to an explanatory theory (from *Figure 2a* to *Figure 2b*) is visualis by replacing the black box with the contents of the other side of the moc Such a theory should be based on the present assumptions of the field. Th assumptions are:

(1)   Perception of the content of a document is primarily based on t presence of and proximity relations among words in text (cf. Giulia₁ 1965, page 26; Harter, 1975, page 199).
(2)   Perceived relations among concepts can be observed in text in t relations among the words representing the concepts (Giuliano, 19 page 223; 1965, page 26)*.
(3)   All occurrences of the same word in a document (after stemmi₁ represent the same concept (Harter, 1975, page 201).
(4)   A fundamental relationship among concepts, texts and words is similar (*see* Woelfel, 1974; Koll, 1979b, pages 35–37).

It has been pointed out by Bar-Hillel (1975) and Robertson (1977) th progress in IR hinges on the ability to understand the content of text. To ma researchers or philosophers of IR, this means 'understanding natu language'. However, for document retrieval or many behavioural choices o a very limited kind of understanding is needed. It is only necessary to say whi of two pairs of information items is more similar. That is, an IR system dc not have to understand much of the meanings of the concepts 'psycholog 'cognition' and 'turkey', only that the first two are the most closely relat This type of meaning (that is, the relative similarity of concepts) does r require sophisticated linguistic theory, but can be observed in the patterns the tokens representing those documents in text.

(5)   Queries are processed over the index language in the same manner documents.

---

* These two assumptions embody the idea that semantic and statistical properties of text strongly related. This point has a long history of discussion and research (*see*, for example, Luhn, 1957; Katter, 1967; Kim and Kim, 1977).

6.1.4.2   Including the user

Including a model of the user's concept relations* in a theory requires a methodology for modelling those relations. The literature on theory in IR has advanced the idea that models of how humans organise information should be examined for potential contribution. Some proposals for where these models should come from and what they should look like have been made by Paisley (1968), Belkin and Robertson (1976) and Smith (1976).

While any model of concept relations provides additional information above that now used in many IR systems, content restrictions and simplicity can be used to narrow the field. The approach adopted here limits concept relations to similarity. Therefore, the more complex models (such as predicate calculus, production systems, complex semantic nets and frames: see Smith, 1976) need not be considered, since similarity can be modelled more simply by either spatial or simple network models.

Arguments can be made for both model forms (see Craig, 1975). The spatial model was selected as more appropriate for use in implementing the SAM model (Koll, 1979b). In particular, Woelfel's GALILEO spatial model of cognition and communication effects was used†. The reliability and validity of the GALILEO data-gathering methodology has been demonstrated (Wisan, 1972; Gordon, 1976; Craig, 1977; Gillham and Woelfel, 1977; Koll, 1979b).

6.1.4.3   A theory

The model shown in *Figure 6.2(b)*, incorporating state of the art assumptions and a component for the users' concept relations, constitutes a testable theory of IR (or, more precisely, of human SRD judgements). Under the SAM model, the steps taken by the system are hypothesised to be the same as those intervening stages of the user. Research under this model searches for congruence between the corresponding stages of system and user. *Figure 6.2(a)* implies that with all other factors constant (that is, the retrieval function and the method of describing documents and information requirements over the concept space), the quality of the match between the system's and users' SRDs

* The terms 'concept', 'concept relations' and 'concept space' are used throughout this work. 'Concept' is defined as the reference of any symbol or collection of symbols, such as a word, stem, title or abstract. A concept may be operationally defined by the model used to manipulate it (for example, as a point in multidimensional space). 'Concept relations' are limited here to similarity: a single number representing the similarity between any two concepts. The operational definition of the users' (perceived) relation between two concepts is the mean judgement of the difference between the concepts reported by users on a questionnaire. An IR system's concept relations are operationally defined by the method of computing them from their (co-)occurrences in the database. 'Concept space' is a generic term for models of concept relations. It does not have to take the form of a vector space. It could be described by networks, sets or other formalisms.

† The GALILEO space is a model used to describe and predict communication and behavioural phenomena (see Woelfel, 1971; Saltiel and Woelfel, 1975; Gillham and Woelfel, 1977). The space is created by multi-dimensional scaling of pairwise, ratio-level dissimilarity judgements of concepts made by human subjects. The GALILEO space is designed to deal with changing relationships over time. The linear force aggregation theory (LFAT) has been proposed as a law of motion through the space (Woelfel, 1971; Saltiel and Woelfel, 1975). LFAT states that an attitude (or behaviour, etc.), defined as a vector in space, converges on the mean of the attitudes encountered by the individual. It has shown good predictive validity for behaviours such as voting preferences, occupational choice, smoking rates and belief change (Danes, Hunter and Woelfel, 1978).

is dependent upon the agreement between the system's and users' conc
similarities.

This dependence becomes a crucial statement of the SAM theory, since it
claim not made explicitly by the SAT model. The SAM theory also implie
preference for using co-occurrence data over treating terms independen
This preference arises from the observation that automatically genera
concept relations which consider term dependencies should agree more w
the users' concept relations than relations based on an independe:
assumption (where all concepts are equally dissimilar).

These two points clearly differentiate the SAM and SAT models. The S/
theory (1) claims that the quality of a system's SRD predictions is depend
on the agreement of its concept similarities with those of the user and
implies that a system using dependency information should predict S]
better than a system not using that information. The SAT model makes
statement on either point. Thus, a critical test can be defined on the S/
theory's added statements.

6.2   A test of the theory

6.2.1   Method

An empirical study was conducted to test the validity of the statements wh
differentiate the SAT and SAM models. The essential question is whether
ability to predict SRD (the match between boxes Nos. 4 and 6 in *Figure 6.2l*
dependent on the correspondence between concept similarities (boxes No.
and 5).

On the basis of the SAM theory, it may be hypothesised that
improvement in the agreement between an IR system's and users' conc
similarities improves the system's predictions of SRD.

If the ability to predict SRD is dependent upon the concept similarity mat
and a system utilising term dependencies improves the concept similar
match, then the system using the dependency information should prodi
more accurate SRD predictions. Thus, it may be hypothesised that using te
dependencies improves an IR system's predictions of SRD and that
improvement is due to the increased agreement between system's and use
concept similarities.

The experiment consisted of comparing the SRDs predicted by an IR syst
using independent, dependent and perceived concept similarities, with 1
SRDs reported directly by a user sample. The experiment required:

(1)   A computer program to predict SRD as would be done by three
      systems differing only in that they use independent, dependent a
      perceived concept similarities, respectively.
(2)   Three queries and twenty documents that had been retrieved for each i
      prior study (McGill, Koll and Noreault, 1979).
(3)   A questionnaire, filled out by 24 subjects, on which they reported th
      perceived dissimilarities between concepts and between requests a
      documents (i.e. SRD).

The outputs of three retrieval systems were simulated:

SYSTEM I   (Independent). Predicts SRD based on the independence

sumption. There are no degrees of relationship among terms; they are all equally dissimilar.

SYSTEM D   (Dependent). Predicts SRD based on concept similarities which vary according to their level of co-occurrence in the text of the database.

SYSTEM P   (Perceived). Predicts SRD based on perceived concept similarities as obtained from the user population.

All three systems may be described by the same formula (6.1), which has a single between-systems variable to represent the differences in concept similarities.

$$
SR_jD_kX = \sqrt{\left\{ \sum_{i=1}^{M} \left[\frac{d_{ik}(c_{ijx} - d_{ik} \cdot q_{ij}(c_{ijx}) + q_{ij})}{|D_k|} - \frac{q_{ij}(p_{ikx} - d_{ik} \cdot q_{ij}(p_{ikx}) + d_{ik})}{|Q_j|} \right]^2 \right\}}
$$
(6.1)

where, for an IR system X, SR_jD_k = predicted similarity between document k (D_k) and request j (R_j) (is actually dissimilarity; SR_jD_k ranges from 0, if items are identical, up through the largest distance possible in the space); Q_j = query for request j; M = number of terms in database; n = number of terms in document k; m = number of terms in query j; d_{ik} = 1 if term i occurs in document k, 0 otherwise; q_{ij} = 1 if term i occurs in query j, 0 otherwise; c_{ijx} = distance of term i from request j in concept space of system X; p_{ikx} = distance of term i from document k in concept space of system X; $|D_k|$ = length of document k

$$
= \sqrt{\left(\sum_{i=1}^{M} d_{ik}^2; |Q_j| \right)} = \text{length of query } j = \sqrt{\left(\sum_{i=1}^{M} q_{ij}^2 \right)}
$$

The differences between systems are reflected in the term-query and term-document distances (c and p in formula 6.1). *Table 6.1* shows what happens to the distance between a document and a query as each term is considered by the different systems.

For System I, the c and p values are a constant: 1. In this way the SRDs produced by formula (6.1) are equivalent to those that would be produced by computing the cosine correlation between the query and the document vectors and converting to distance.

For System D, the distance from any term, x, to information items (queries or documents) which do not contain that term varies according to the degree to which term x co-occurs with the terms contained in the information item. The distances are computed from a term–term co-occurrence matrix using formulas (6.2) and (6.3).

$$
c_{ijD} = \sqrt{\left[2 - 2/m \left(\sum_{g=1}^{m} r(x_iy_g) \right) \right]}
$$
(6.2)

$$
p_{ikD} = \sqrt{\left[2 - 2/n \left(\sum_{h=1}^{n} r(x_iy_h) \right) \right]}
$$
(6.3)

where x_i = any term i; y_g = gth term in query J; y_h = hth term in document k; m = No. of query terms; n = No. of terms in document k; $r(x_iy_z)$ = correlation between terms x_i and y_z.

System P embodies a perfect concept space match, since it uses the users' perceived concept similarities. The difference between System D and System P

TABLE 6.1. Differences among SRDI, SRDD and SRDP

	SRDI	SRDD	SRDP
Concept similarity structure	Independent. All term–information item distances are equal.	Dependent. Term–information item distance is inversely related to degree of co-occurrence of term and terms in information item.	Perceived. Term–information item distance is as rep by users on GALILEO-type questionnaire.
Range of c_{ij} and p_{ik}	min. $= 1$; max. $= 1$	min. $= 0$; max. $= \sqrt{2}$	min. $= 0$; max. $= i$ $(= 100)^*$
Range of SRD	min. $= 0$; max. $= \sqrt{2}$	min. $= 0$; max. $= \sqrt{2}$	min. $= 0$; max. $= i$ $(= 15)^*$

Increase in SRD^2 when term i is summed over, when:

$d_{ik} =$	$q_{ij} =$			
0	0	0	0	0
1	1	$\left(\dfrac{1}{\sqrt{m}} - \dfrac{1}{\sqrt{n}}\right)^2$	$\left(\dfrac{1}{\sqrt{m}} - \dfrac{1}{\sqrt{n}}\right)^2$	$\left(\dfrac{1}{\sqrt{m}} - \dfrac{1}{\sqrt{n}}\right)^2$
0	1	$1/m$	$\dfrac{|i|^2 + |k|^2 - 2|i||k|\cos(ik)^\dagger}{m}$	$\dfrac{\text{perceived dist. }(i,k)}{m}$
1	0	$1/n$	$\dfrac{|i|^2 + |j|^2 - 2|i||j|\cos(ij)^\dagger}{n}$	$\dfrac{\text{perceived dist.}(i,j)}{n}$

* Perceived distances are measured on open-ended scale. While theoretically there is no maximum, in practice maxi values tend to be as indicated.

† i, j and k here refer to vectors representing the term i, document k and query j. '|' indicates vector length. The vectors cons presence/absence data of (1) the term i over the set of documents in the database, (2) the average of the terms in document k over rest of the database and (3) the average of the terms in query j over the database.

is the origin of the term–term similarities. In System P the distances obtained by asking a sample of the user population to judge the dissimilarit between concepts. The questionnaire is based on the GALILEO format (Gordon, 1976; Koll, 1979b).

These three sets of predicted SRDs (SRDI, SRDD and SRDP) evaluated by their correspondence to the dependent variable: observed SI (SRDO). SRDO is obtained via the same questionnaire and method as the and p_{ikP} values. Respondents are asked to judge the similarity between retrieved document and the query. SRDO for a given document and query defined as the mean judgement of the respondents.

The statistical hypotheses to be tested are:

H1: $\rho(PO) > \rho(IO)$
H2: $\rho(PO) > \rho(DO)$

H3: $\rho(DO) > \rho(IO)$
H4: $\rho(DO \cdot P) = \rho(IO \cdot P)$

where I, D and P are the SRDs predicted by the systems using independent, dependent and perceived concept relations, O is the SRDs reported by users, $\rho(XO)$ is the correlation between predicted SRDXs and SRDO, and $\rho(XO \cdot P)$ is the partial correlation between the SRDs predicted by System X and the SRDs reported by users, with the effect of the SRDs predicted by System P controlled.

H1 and H2 predict that the SRDs predicted by a system P which has perfect agreement with users' concept similarities will have a higher correlation with, perceived SRDs (SRDO) than will the SRDs predicted by systems I and D, which have less than perfect concept similarity matches. H3 is similar to H1 and H2. We assume that D has better concept similarities agreement than I; therefore, it should have better SRD predictions.

With the constraints discussed above, the SAM theory implies that SRD predictions will improve only to the extent caused by the concept similarities match. It is hypothesised that if between the two systems D and I a difference in ability to predict SRD is found, that difference will be due to a difference in their concept similarities. Therefore, when the effect of the concept similarities match is controlled, the difference in SRD predictions should not remain.

Such a difference is expected to be found between Systems D and I (hypothesis H3). H4 claims that the difference between Systems D and I in ability to predict SRD will not remain when the effect of System P's SRDs is controlled. Controlling for System P's effect controls the effect of a perfect concept similarities match. Any remaining difference between Systems D and I in ability to predict SRD would have to be attributed to factors outside the theory. Together H3 and H4 state that the use of dependency information will improve prediction of SRD, but only because such information improves the concept similarity match.

The Pearson product moment correlation was used to measure the relations among the SRDs, since they are all measured continuously. The testing procedures for correlated rs and partial rs are discussed in Hotelling (1940) and McNemar (1969). Power analysis (see Cohen, 1969) was used to determine sample size. An n of 60 documents was chosen to balance the risks of Type 1 and Type 2 errors (see Koll, 1979b).

It should be noted that in this analysis the correlations are between lists of SRDs, combined over several queries. Secondary analyses of individual queries were also done (see Koll, 1979b). The results of those analyses are referred to briefly.

6.2.2   Results

The four hypotheses and test results are shown in *Table 6.2*. All four null hypotheses are rejected at the .1 level. The statistical conclusions are: H1-A $(\rho(PO) > \rho(IO))$; H2-A $(\rho(PO) > \rho(DO))$; H3-A $(\rho(DO) > \rho(IO))$; and H4-A $(\rho(DO \cdot P) > \rho(IO \cdot P))$. The first three conclusions support the SAM theory by confirming its predictions that estimates of user perceptions of documents, as made by current systems, can be improved by using term dependency information and more so by using perceived concept similarities.

TABLE 6.2 Initial hypothesis testing

Hypothesis		Conclude
H1-0: $\rho(PO) = \rho(IO)$	$r(PO) = \cdot 63$; $r(IO) = \cdot 45$; $r(PI) = \cdot 54$	
H1-A: $\rho(PO) > \rho(IO)$	$t = 1 \cdot 850 > t^*$ (df = 57) = 1·296	$\rho(PO) > \rho(IO)$
H2-0: $\rho(PO) = \rho(DO)$	$r(PO) = \cdot 63$; $r(DO) = \cdot 48$; $r(PD) = \cdot 55$	
H2-A: $\rho(PO) > \rho(DO)$	$t = 1 \cdot 71 > t^*$ (df = 57) = 1·296	$\rho(PO) > \rho(DO)$
H3-0: $\rho(DO) = \rho(IO)$	$r(DO) = \cdot 48$; $r(IO) = \cdot 45$; $r(DI) = \cdot 99$	
H3-A: $\rho(DO) > \rho(IO)$	$t = 1 \cdot 865 > t^*$ (df = 57) = 1·296	$\rho(DO) > \rho(IO)$
H4-0: $\rho(DO \cdot P) = \rho(IO \cdot P)$	$r(DO \cdot P) = \cdot 20$; $r(IO \cdot P) = \cdot 17$; $r(DI \cdot P) = \cdot 99$	
H4-A: $\rho(DO \cdot P) > \rho(IO \cdot P)$	$t = 1 \cdot 529 > t^*$ (df = 56) = 1·296	$\rho(DO \cdot P) > \rho(IO \cdot P)$

$n = 60$; $a = \cdot 1$, one-sided.

When the effect of SRDP is controlled, both SRDD and SRDI dro₁ significantly in their ability to predict SRDO. This supports the claim tha perceived concept similarities are an important component of good SRΓ predictions. However, the difference between the success rates of SRDD an< SRDI is virtually unchanged and still statistically significant. This departur₁ from the hypothesised result can be explained by the fact that the ful relationship between SRDP and SRDO was not captured by the linear model With transformed c_{ij} and p_{ik}, the correlation between SRDP and SRDΓ increases from .55 to .77.

Examining the relationships between SRDP and SRDD, and SRDO, it wa observed that monotonic transformations of c_{ijD} and p_{ikD} and c_{ijP} and p_{ik} produced large improvements in the linear relationships between SRDP an< SRDO and between SRDD and SRDO. Using the square root of both c_{ijP} an< p_{ikP} increased the correlation between SRDP and SRDO from .63 to .66. Fo the correlation between SRDD and SRDO it was observed that a maximun (of about .53) occurred with c_{ijD} and p_{ikD} each raised to powers near 7 or 8

The statistical results from this second analysis are given in *Table 6.3* Statistical conclusions are all as predicted by the theory: H1-A $(\rho(PO) > \rho(IO))$ H2-A $(\rho(PO) > \rho(DO))$, H3-A $(\rho(DO) > \rho(IO))$ and H4-0 $(\rho(DO \cdot P) = \rho(IO \cdot P))$. From this analysis it may be concluded that System P > System Γ > System I, and that System D = System I when the effect of System P i controlled.

The claim that the dependency information is tapping information abou perceived concept relations is also supported by other statistics. Th₁ correlation between SRDD and SRDP is higher than between SRDI an< SRDP $(r(PD) = .77$, $r(PI) = .71$; $t = 2.24$, $p(t > 2.24$, df = 57, one alt.) < .025) Also, the correlation of SRDD, with the effect of SRDI removed, with SRDP i significantly higher than that of SRDI, with SRDD removed $(r(P(D \cdot I) = .31$ $r(P(I \cdot D) = -.05)$. In other words, when one examines the non-share< components of SRDD and SRDI, SRDD is much more strongly related t₁ SRDP. Further, multiple regression analysis (*see* Koll, 1979b) indicates tha there is no significant correlation between either SRDI or SRDD and SRDC after SRDP is accounted for. In fact, the independent and dependent system add only .01 to the multiple correlation coefficient, from .66 to .67.

TABLE 6.3. Transformed data hypothesis testing

Hypothesis		Conclude
H1-0: $\rho(PO) = \rho(IO)$	$r(PO) = \cdot66$; $r(IO) = \cdot45$; $r(PI) = \cdot71$	
H1-A: $\rho(PO) > \rho(IO)$	$t = 2\cdot77 > t^*$ (df = 57) = 1·296	$\rho(PO) > \rho(IO)$
H2-0: $\rho(PO) = \rho(DO)$	$r(PO) = \cdot66$; $r(DO) = \cdot53$; $r(PD) = \cdot77$	
H2-A: $\rho(PO) > \rho(DO)$	$t = 1\cdot928 > t^*$ (df = 57) = 1·296	$\rho(PO) > \rho(DO)$
H3-0: $\rho(DO) = \rho(IO)$	$r(DO) = \cdot53$; $r(IO) = \cdot45$; $r(DI) = \cdot94$	
H3-A: $\rho(DO) > \rho(IO)$	$t = 1\cdot877 > t^*$ (df = 57) = 1·296	$\rho(DO) > \rho(IO)$
H4-0: $\rho(DO\cdot P) = \rho(IO\cdot P)$	$r(DO\cdot P) = \cdot04$; $r(IO\cdot P) = -\cdot03$; $r(DI\cdot P) = \cdot88$	
H4-A: $\rho(DO\cdot P) > \rho(IO\cdot P)$	$t = 1\cdot07 < t^*$ (df = 56) = 1·296	$\rho(DO\cdot P) = \rho(IO\cdot P)$

$n = 60$; $a = .1$, one-sided.

TABLE 6.4. Accounting for SRDO's variance (transformed data)

Variables	r (over all Qs)	90% confidence interval
SRDI and SRDO	·45	$\cdot26 < \rho < \cdot61$
SRDD and SRDO	·53	$\cdot36 < \rho < \cdot67$
SRDP and SRDO	·66	$\cdot52 < \rho < \cdot77$
SRDD·P and SRDO·P	·04	$-\cdot18 < \rho < \cdot25$
SRDI·P and SRDO·P	$-\cdot03$	$-\cdot25 < \rho < \cdot19$

| Variables | Difference between variances' r-squared with SRDO | Increase in % variance explained | | |
		mean (all Qs)	min. (one Q)	max. (one Q)
SRDP − SRDI	$\cdot44 - \cdot20 = \cdot23$	23%	8%	47%
SRDP − SRDD	$\cdot44 - \cdot28 = \cdot15$	15%	−9%	44%
SRDD − SRDI	$\cdot28 - \cdot20 = \cdot08$	8%	3%	17%
SRDD·P − SRDI·P	$\cdot00 - \cdot00 = \cdot00$	0%	0%	16%

Variables marked X·Y = residuals from X regressed on Y.

Confidence intervals (for transformed data) for the simple and partial correlations are included in *Table 6.4*. Also given are the expected values for increases in the percentage of SRDO's variance accounted for (that is, incremental r-squared). These values provide a straightforward indication of the effect of dependent and perceived concept similarities on the predictability of SRDO. The baseline prediction by binary vectors using the cosine correlation can account for 20 per cent of SRDO's variance. Use of perceived concept relations can increase this by another 23 per cent to 44 per cent. Of that increase, 8 per cent could have been achieved by using dependency relations as an estimate of perceived relations.

The conclusions to be drawn from these experimental data are generally supportive of the hypotheses. The evidence is strongly supportive of the first three hypotheses: System P > System D > System I in predicting SRDO. H-4

is also supported, but less strongly. Support is stronger at the highe aggregation level (over queries) than when individual queries are analyse separately (*see* Koll, 1979b). The conclusions should also be tempered witl conventional caution about generalising from 3 queries, 60 documents and 2 users. General conclusions are:

(1) Improving the agreement between an IR system's and its users' concep similarities improves the system's predictions of SRD. An additional 2 per cent of SRD's variance may be accounted for by using perceive instead of independent concept similarities.

(2) Some of this gain may also be attained by using dependent concep similarities. The expected loss in SRD predictability due to the inde pendence assumption is at least 8 per cent.

(3) While it is possible that dependency information taps factors relating t SRD other than those of perceived concept similarities, it seems that th improvement caused by dependency information is due to the fact tha dependent concept similarities approximate perceived concep similarities.

(4) On the basis of the support for statements made by the SAM model bu not by the SAT model, the SAM model should be accepted as a mor accurate description of information retrieval.

6.3   Implications

6.3.1   Introduction

Essentially, what has been learned is: (1) the ability to predict a use population's SRDs is (at least partially) dependent on the ability to predic their inter-concept similarities; and (2) using co-occurrence information is on (at least partially) effective way of doing that. This knowledge has severa implications for future IR research.

6.3.2   Improved perceived model

A research programme on concept space formation might be useful. It coulc address the issue of what model(s) of the perceived concept space, as well a: what methods of finding document locations from concepts locations, are bes for predicting SRD. Alternatives include greater exploration of the GALILEC space as well as other formalisms. The laws of motion through the GALILEC space need exploration. The Linear Force Aggregation Theory (whicl indicates that an information item is located at the centroid of its parts) wa: used with some success in this study. Other laws of motion are possible, botl for aggregating from simple to complex concepts and for aggregating ove: people.

For example, there is some evidence that the geometric mean is mor appropriate than the arithmetic mean for pooling perceived magnitude: (Stevens, 1975). Also, different ranking algorithms (see McGill,Koll anc Noreault, 1979) from that used here would translate into different ways o locating documents and queries (other than the centroid method) and ways o measuring similarity other than Euclidean distance. It would also b interesting to see how well the proposed GALILEO space laws abou

concepts' resistance to change (based on accumulated mass of messages: *see* Saltiel and Woelfel, 1975) hold up through database updating.

Different similarity assessment procedures may indicate a different model form. It would be instructive to compare, for example, a network model of perceived relations against the GALILEO space in terms of SRD predictions.

6.3.3   Approximating perceived space

Given the improvement shown by the perceived system over the independent system, but the effort required to construct the users' concept space via questionnaire, a research programme aimed at approximating that space by automatic analysis of the database could be valuable. Even though co-occurrence data may not be as powerful as users' perceptions (for example, 8 per cent versus 23 per cent increment in r-squared) the savings involved in working completely automatically, instead of requiring manual (user) effort, make the use of co-occurrence data attractive.

In the present study the automatically generated space achieved only one-third of the increase in performance of the perceived space. This leaves much room for improvement.

Unfortunately, in this study economic limitations prohibited a direct comparison of inter-concept similarities as perceived and as observed in co-occurrence data. A simple correlational analysis would involve randomly selected concept-pairs and a comparison of the two corresponding sets of similarity values.

Maximising that correspondence would, according to this theory, maximise SRD predictability. Independent variables that could help improve the correspondence include term weighting schemes (attending to factors such as collection frequency), similarity measures, stemming routines, and methods of defining co-occurrence (for example, same document, same sentence, n-word window).

6.3.4   Using concept relations

Slightly different from the issue of maximising the concept similarities match is a more pragmatic question of how IR practitioners can utilise concept relations (either perceived or empirically derived).

In the 1960s and early 1970s a number of methods of using concept relations were proposed (*see*, for example, Doyle, 1961; Stiles, 1971; Giuliano, 1963; Dennis, 1965; Switzer, 1965; Curtice, 1966; Katter, 1967; Lesk, 1969; Cagan, 1970; Minker, Wilson and Zimmerman, 1972). Recently, there has been a revival of interest in this area. Methods of using concept relations have been proposed or experimentally implemented by Attar and Frankel (1977), Bookstein and Kraft (1977), Harper and van Rijsbergen (1978), Yu and his associates (*see* Yu, Luk and Siu, 1978; Raghaven and Yu, 1979), Belkin and Oddy (1979), Croft (1979), Doszkocs (1979), Koll (1979a, 1979b), and PAR Corporation (Laudauer and Mah, 1979). Many different approaches to using concept relations are represented among these works, including multi-dimensional scaling, heuristic space construction, document clustering, decision theory and straight matrix multiplication. These approaches should be

compared and analysed towards the goal of designing an IR system capable of making optimal use of concept relations.

6.3.5   Individual level

For substantive and methodological reasons the research reported here was conducted at an aggregate level. As a follow-up it would be interesting to see whether the results obtained here hold for the individual. There are two obvious research projects: (1) using the concept space constructed at the aggregate level, to predict a single user's SRDs; and (2) using an individual's own concept relations, to build a space in which to predict his SRDs.

The aggregate-to-individual level analysis is more interesting. This approach has potential for short-term gain, since IR systems currently act in the aggregate-to-individual mode. Such studies would differ from the present study only in that the units of analysis would be individual queries rather than documents aggregated across queries.

The individual-to-individual analysis could lead to improvement in IR only if it becomes technically and economically feasible to individualise the system for each user. Major alteration of the concept space for each individual is probably not a practical approach to IR. Some individualisation, in the form of relevance feedback or minor modification of the space based on limited user input, could be practical and useful.

There are benefits to studying both levels. One interesting comparison would be of the retrieval effectiveness of the population's concept space versus that of the individual's own concept space. It is conceivable that the population is better able to structure the field in which an individual is seeking information than that person can himself.

References

ATTAR, R. and FRANKEL, S. (1977). 'Local feedback in full text retrieval systems', *Journal of the ACM*, **24**, 397–417

BAR-HILLEL, Y. (1975). 'The impact of the essentially pragmatic character of the natural languages on linguistic information processing', in DEBONS, A. and CAMERON, W. J. (Eds), *Perspectives in Information Science* (NATO Advanced Study Institute Proceedings, Aberystwyth, Wales, 1973), pp. 297–305, Noordhoff, The Hague

BELKIN, N. J. and ODDY, R. N. (1979). *Design Study for an Anomalous State of Knowledge Based Information Retrieval System*, Final Report to British Library Research and Development Department, Grant S1/SG/09

BELKIN, N. J. and ROBERTSON, S. E. (1976). 'Information science and the phenomenon of information', *Journal of the ASIS*, **27**, 197–204

BOOKSTEIN, A. and COOPER, W. (1976). 'A general mathematical model for information retrieval systems', *Library Quarterly*, **46**, 153–167

BOOKSTEIN, A. and KRAFT, D. H. (1977). 'Operations research applied to document indexing and retrieval decisions', *Journal of the ACM*, **24**, 418–427

BORING, E. (1953). 'The role of theory in experimental psychology', *American Journal of Psychology*, **LXVI**, 169–184

CAGAN, C. (1970). 'A highly associative information retrieval system', *Journal of the ASIS*, **21**, 330–337

COHEN, J. (1969). *Statistical Power Analysis for the Behavioral Sciences*, Academic Press, New York

CRAIG, R. T. (1975). 'Models of cognition, models of messages and theories of communication effects: spatial and network paradigms', paper presented at ICA annual meeting, Chicago

CRAIG, R. T. (1977). 'Limiting the scope of the spatial model of communication effects', *Human Communication Research*, **3**, 309–32

CROFT, W. B. (1979). *Organizing and Searching Large Files*, Ph.D. Thesis, University of Cambridge

CURTICE, R. M. (1966). 'Experiments in associative retrieval', in *Proceedings of American Documentation, Institute Annual Meeting*, Santa Monica, California, pp. 373–383

DANES, J. E., HUNTER, J. E. and WOELFEL, J. (1978). 'Mass communication and belief change: a test of three mathematical models', *Human Communication Research*, **4**, 243–252

DENNIS, S. (1965). 'The construction of a thesaurus automatically from a sample of text', in STEVENS, M. E. *et al.* (Eds), *Statistical Association Methods for Mechanical Documentation*, Symposium Proceedings, Washington, D.C., 1964 (NBS Miscellaneous Publication No. 269), pp. 61–148

DOSZKOCS, T. E. (1979). *AID: An Associative Interactive Dictionary for Online Bibliographic Searching*, Ph.D. Thesis, University of Maryland

DOYLE, L. B. (1961). 'Semantic road maps for literature searches', *Journal of the ACM*, **8**, 553–578

GILLHAM, J. and WOELFEL, J. (1977). 'The GALILEO system of measurement: preliminary evidence for precision, stability, and equivalence to traditional measures', *Human Communication Research*, **3**, 222–234

GIULIANO, V. E. (1963). 'Analog networks for word associations', *IEEE Transactions on Military Electronics*, Mil, 221–225

GIULIANO, V. E. (1965). 'The interpretation of word associations', in STEVENS, M. E. *et al.* (Eds), *Statistical Association Methods for Mechanical Documentation*, Symposium Proceedings, Washington, D.C., 1964 (NBS Miscellaneous Publication No. 269)

GORDON, T. F. (1976). 'Subject abilities to use metric MDS: effects of varying the criterion pair', paper presented at Education for Journalism

HARPER, D. J. and VAN RIJSBERGEN, C. J. (1978). 'An evaluation of feedback in document retrieval using co-occurrence data', *Journal of Documentation*, **34**, 189–216

HARTER, S. (1975). 'A probabilistic approach to keyword indexing: Part 1', *Journal of the ASIS*, **26**, 197–206

HOTELLING, H. (1940). 'The selection of variates for use in prediction with some comments on the general problem of nuisance parameters', *Annals of Math. Stat.* **11**, 271–283

KATTER, R. V. (1967). *Study of Document Representations; Multidimensional Scaling of Index Terms*, TM-3627, Systems Development Corporation, Santa Monica, California

KIM, C. and KIM, S. D. (1977). 'Consensus vs frequency: an empirical

investigation of the theories for identifying descriptors in designing retrieval thesauri', *Information Processing and Management*, **13**, 253–258

KOLL, M. B. (1979a). 'WEIRD: an approach to concept-based information retrieval', *ACM SIGIR FORUM*, **XIII**, 32–50

KOLL, M. B. (1978b). *The Concept Space in Information Retrieval Systems as a Model of Human Concept Relations*, Ph.D. Thesis, School of Information Studies, Syracuse University, Syracuse, N.Y.

LAUDAUER, C. and MAH, C. (1979). 'Message extraction through estimated relevance', *ACM SIGIR FORUM*, **XIV**, 64–70

LESK, M. E. (1969). 'Word–word associations in document retrieval', *American Documentation*, **20**, 27–38

LUHN, H. P. (1957). 'A statistical approach to mechanized encoding and searching of literary information', *IBM Journal of Research and Development*, **1**, 309–317

McGILL, M. J., KOLL, M. B. and NOREAULT, T. (1979). *Evaluation of Factors Affecting Document Ranking by Information Retrieval Systems: Final Report*. School of Information Studies, Syracuse University, Syracuse, N.Y.

McNEMAR, Q. (1969). *Psychological Statistics*, Wiley, New York

MARON, M. E. (1977). 'On indexing retrieval and the meaning of about', *Journal of the ASIS*, **28**, 38–43

MARON, M. E. and KUHNS, J. L. (1960). 'On relevance, probabilistic indexing and information retrieval', *Journal of the ACM*, **7**, 216–244

MINKER, J., WILSON, G. A. and ZIMMERMAN, B. H. (1972). 'An evaluation of query expansion by the addition of clustered terms for document retrieval system', *Information Storage and Retrieval*, **8**, 329–348

PAISLEY, W. J. (1968). *As We May Think Information Systems Do Not*, speech to American Psychological Association, San Francisco, Califonia, ED037095

POPPER, K. R. (1965). *Objective Knowledge: An Evolutionary Approach*, Oxford University Press, London

RAGHAVEN, V. V. and YU, C. T. (1979). 'Experiments on the determination of the relationships between terms', *ACM Transactions of Database Systems*, **4**, 240–260

ROBERTSON, S. E. (1977). 'Theories and models in information retrieval', *Journal of Documentation*, **33**, 126–148

SALTIEL, J. and WOELFEL, J. (1975). 'Interia in cognitive processes: the role of accumulated information in attitude change', *Human Communication Research*, **1**, 333–334

SMITH, L. C. (1976). 'Artificial intelligence in information retrieval', *Information Processing and Management*, **12**, 189–222

STEVENS, S. S. (1975). *Psychophysics*, Wiley, New York

STILES, H. E. (1961). 'The association factor in information retrieval', *Journal of the ACM*, **8**, 271–279

SWANSON, D. R. (1977). 'Information retrieval as a trial and error process', *Library Quarterly*, **47**, 128–148

SWITZER, P. (1965). 'Vector images in information retrieval', in STEVENS, M. E. *et al.* (Eds), *Statistical Association Methods for Mechanical Documentation*, Symposium Proceedings, Washington, D.C., 1964 (NBS Miscellaneous Publication), pp. 163–171

WISAN, G. (1972). *The Construction of a Continuous Manifold for Social Objects: A Prolegomenous Study*, Ph.D. Thesis, Sociology Department, University of Illinois

WOELFEL, J. (1974). 'Metric measurement of cultural processes', paper presented at Annual Meeting of Speech Communication Association, Chicago, Illinois

WOELFEL, J. (1971). *Sociology and Science*, unpublished manuscript, Department of Communication, Michigan State University

YU, C. T., LUK, W. S. and SIU, M. K. (1978). 'On the estimation of the number of desired records with respect to a given query', *ACM Transactions on Database Systems*, 3, 41–56