The Use of Anaphoric Resolution for Document Description in Information Retrieval

Susan Bonzi and Elizabeth Liddy

School of Information Studies
Syracuse University

Abstract

This study investigated two hypotheses concerning the use of anaphors in information retrieval. The first hypothesis, that anaphors tend to refer to integral concepts rather than to peripheral concepts, was well supported. Two samples of documents, one in psychology and the other in computer science, were examined by subject experts who judged the centrality of phrases which were referred to anaphorically. The second hypothesis, that various term weighting schemes are affected differently by anaphoric resolution, was also well supported. It was found that schemes which incorporate document length into the calculations produce much smaller increases in term weights for terms occurring in anaphoric resolutions than do those which do not consider document length. It is concluded that although anaphoric resolution has potential for better representing the "aboutness" of a document, care must be taken in choosing both the anaphoric classes to be resolved and the term weighting schemes to be used in measuring a document's topicality.

I. Introduction

For the past several years there has been a tendency in the field of information retrieval to think that the most sensible direction for our systems to take is toward the use of naturally occurring text - both in the information source and in the information query. In addition, many of us have come to believe that the statistical techniques with which our field evolved and which are still used when dealing with natural language texts have reached the limits of their usefulness in achieving greater correlation between users' needs as expressed in their queries and the desired information as found in the documents or their representations. Information retrieval systems which allow users to state their needs directly to the system without the mediation of a structured query language and which contain useful representations of the semantic content of documents (as intended but not necessarily explicitly stated by their authors) will require natural language processing techniques of greater complexity than those currently available.

One of the greatest difficulties encountered in accurately representing natural language texts is the interpretation of anaphoric references. An anaphor is an abbreviated subsequent reference to a concept mentioned earlier in the text. Pronouns are the most prominent, but not the only types of anaphors. For example, the phrase "an online computerized

information retrieval system" may later be anaphorically referred to as "this system" where "this" functions as an anaphor lexically replacing "online computerized information retrieval". Humans correctly process the anaphor "this" as "online computerized information retrieval" for use in their mental representation of the meaning of the phrase. However, this facility is not fully developed in most natural language processing systems nor in any information retrieval systems of which we are aware.

The need for appropriate handling of anaphoric references in information retrieval systems has long been recognized [O'CONNOR73; PAICE81; SALTON83] but empirical investigation of the phenomenon and its impact on information retrieval systems has only recently begun [KATZER86b; LIDDY87; PAICE87]. [O'CONNOR73] indicated that his results in developing a question-answering system would have been much improved had his system been able to recognize and resolve anaphoric references in the answer-providing texts. Early work by [PAICE81] on automatic extracting suggested that more cohesive extracts were possible if anaphoric references were recognized and the fuller text to which they referred was included. In a text on information retrieval [SALTON83], the resolution of anaphoric references is characterized as one of the more difficult problems in language analysis which has been commonly ignored in our field's attempts at developing natural language question-answering systems.

Paice and Husk, continuing the investigation into automatic methods for generating document abstracts [PAICE87], conducted a very detailed and extensive analysis of the use of the pronoun "it" in 1255 instances in a test collection of technical English texts. They then developed computer programs for detecting whether each occurrence of "it" was anaphoric or not with a success rate of 92.2%. These promising results suggest that extension of their techniques to other anaphors may be useful to any system which must process and represent the meaning of naturally occurring text.

The impetus behind the present investigation was an extensive experimental study of the impact of anaphoric references on information retrieval which was carried out at Syracuse University [KATZER86a]. The first phase of the project was concerned with developing linguistic rules for deciding when occurrences of each of 142 potentially anaphoric terms, categorized in 10 linguistic classes (Figure 1), are in fact functioning anaphors. These rules were based on a study of 600 documents (titles and abstracts) and can form the basis for an automatic procedure to recognize anaphoric terms in bibliographic databases.

Figure 1

	ANAPHOR CLASS	EXAMPLES
1.	Central Pronouns	he, his, it
2.	Nominal Demonstratives	this, these, those
3.	Relative Pronouns	who, which, where
4.	Nominal Substitutes	above, former, none
5.	Pro-verb	do
6.	Indefinite Pronouns	any, each, many
7.	Pro-adjectives	another, identical
8.	Pro-adverbials	so, such, similarly
9.	Subject references	S, Ss
10.	Definite Article	the

The second phase tested whether replacing all anaphors with their referents, a process known as resolution, would affect term frequencies in such a way as to improve retrieval performance. This experiment was based on the commonly held premise that anaphors are used by authors to avoid repetition and as such, they are likely to represent the more important concepts in a document. It was further hypothesized that resolution of all anaphors in a document would favorably affect term weighting schemes based on within document frequency (WDF) measures used by a retrieval system to determine the order in which documents containing any of the query terms should be presented to the user. In a post-retrieval experiment, all functioning anaphors in documents retrieved in response to 24 queries of two databases were resolved. This process changed the WDF of all those words in the documents that were later referred to by an anaphor. The effect was to make the semantic presence of each such term be appropriately reflected in its surface count. Both unresolved and resolved sets of the documents were ranked according to a variety of term weighting schemes, and two similarity measures were used to compare the rankings of each set to the documents ranked according to users' relevance judgments.

The results of the empirical study were mixed. Resolving anaphors did improve ranking for some queries though all classes of anaphora did not contribute equally to this improvement. There were also instances in which ranking performance decreased when certain classes of anaphors were replaced with their referents. In addition, for many queries there was no effect on predicted relevance by the system when anaphors were resolved. We found these results puzzling, but remained convinced that the basic premise underlying this research was valid, namely that anaphors are used to abbreviate subsequent mentions of the more important concepts in a document.

Although the original study made significant contributions to our knowledge of the extent of the presence of anaphors in documents commonly used in free-text retrieval and of how anaphoric uses might be distinguished from non-anaphoric uses, the results raised two major questions. Firstly, we wanted to investigate the basic question of whether anaphors are used by authors of such documents to refer to concepts integral to the document or whether they are used equally to refer to concepts of only peripheral interest to the topic at hand. Because integral concepts tend to be repeated in a document, and since anaphors are used to avoid repetition, it seems reasonable to assume that anaphors tend to refer to integral concepts more often than to peripheral concepts.

The second hypothesis is that different term weighting schemes will react differently when anaphoric resolutions are introduced. The resolution of an anaphor involves the re-introduction of the entire phrase referred to anaphorically. Resolutions comprised of only one or two words will not affect term weighting schemes which take document length into consideration nearly as much as resolutions which add a large number of terms to the length of the document. We hypothesize that term weighting schemes which do not take document length into consideration will, on the average, reflect the increase in the importance of a term referred to by an anaphor to a greater degree than will term weighting schemes which do not.

II. Centrality

The hypothesis that concepts replaced by anaphors in text tend to be conceptually integral to the text was tested in the following manner. 43 abstracts retrieved by online searches of PsycINFO (psychology) and 50 abstracts retrieved by online searches of INSPEC (computer science) constitute the sample from which the data were taken. The abstracts were reformatted so that each resolution, that is, the phrase which was later

referenced by an anaphor, was underlined in the body of the abstract. The phrase was printed again beneath the abstract. Next to each phrase was a scale from 1 to 5. The number of underlined and repeated phrases per abstract ranged from 1 to 11 in the psychology sample and from 1 to 8 in the computer science sample. (The actual number of anaphors in each abstract may have differed because the same phrase could have been referred to by more than one anaphor or two phrases might replace a single anaphor.) The Appendix provides a sample abstract examined by subject experts.

Five people familiar with the literature of psychology (all doctoral students in the Department of Psychology at Syracuse University) and 5 familiar with the literature of computer science (3 doctoral students and 2 second year master's students in the Computer and Information Science Department at Syracuse University) were offered a $50.00 stipend to read a sample of the abstracts from their respective field and to assign a measure of "aboutness" to each underlined phrase. The judges were asked to read each abstract and then to assign a number from 1, "highly integral," to 5, "highly peripheral," to each underlined phrase. They were not given the reason these particular phrases were selected from the abstract. The samples were constructed so that each abstract was judged by three people and each person judged approximately the same number of phrases (between 100 and 120 each).

Two problems in coding the data were encountered. In general, the subject specialists were able to judge each phrase as a whole, but if they were unable to do so, they were asked to split the phrase where appropriate and to assign a measure of aboutness to each part. An additional complication occurred with a small number of anaphors which were resolved by two or more phrases not adjacent in text. Since the phrases were duplicated at the bottom of the page as they appeared in text, it was possible for one measure to be assigned to one part of the anaphoric reference and another measure to be assigned to another part. In most instances where either problem occurred, the measures were very close to each other on the scale, and in those cases, the average was computed. When there was a wide discrepancy between the numbers assigned by an individual judge, the text was examined. In those instances where one phrase was obviously dominant, the measure of the dominant phrase was chosen. In 5 cases in psychology, however, the discrepancy was not easily resolved. Therefore an average of the numbers was computed.

A total of 280 resolutions in PsycINFO and 212 in INSPEC were judged by the subject experts. A measure of aboutness for each resolution was computed by averaging the three judges' individual measures. It became obvious during the coding of the data that one of the five subject experts in the psychology sample tended to judge phrases somewhat differently from the others (generally higher). Both correlational measures and a raw count of differences greater than two supported the observation. Therefore, the data analysis of the psychology sample will include both all subject experts' judgments and all but the outlier's judgments. There was no clear outlier among those judging the INSPEC sample. However, the data were also analyzed with the person whose judgments were most disparate deleted. Table 1 shows the distribution of scores among the classes of anaphors for all judges.

Table 1

Ten Subject Experts' Judgments of "Aboutness"
of Ten Classes of Anaphors

Class	Highly Integral				Highly Peripheral	Total
Central Pronouns						
INSPEC	55 (.495)	19 (.171)	16 (.144)	10 (.090)	11 (.099)	111
PsycINFO	77 (.397)	32 (.165)	53 (.273)	20 (.103)	12 (.062)	194
Nominal Demonstratives						
INSPEC	52 (.495)	25 (.238)	9 (.086)	12 (.114)	7 (.067)	105
PsycINFO	35 (.432)	26 (.321)	11 (.136)	6 (.074)	3 (.037)	81
Relative Pronouns						
INSPEC	28 (.311)	22 (.244)	13 (.144)	8 (.089)	19 (.211)	90
PsycINFO	49 (.272)	24 (.133)	34 (.383)	30 (.167)	43 (.239)	180
Nominal Substitutes						
INSPEC	16 (.333)	16 (.333)	7 (.144)	1 (.021)	8 (.167)	48
PsycINFO	13 (.542)	1 (.042)	7 (.292)	3 (.125)	0 (.000)	24
Proverbials						
INSPEC	0 (.000)	0 (.000)	0 (.000)	0 (.000)	0 (.000)	0
PsycINFO	4 (.333)	0 (.000)	1 (.083)	3 (.250)	4 (.333)	12
Indefinites						
INSPEC	6 (.250)	4 (.167)	4 (.167)	2 (.083)	8 (.333)	24
PsycINFO	25 (.439)	10 (.175)	10 (.175)	7 (.123)	5 (.088)	57
Adjectives						
INSPEC	2 (.222)	2 (.222)	3 (.333)	0 (.000)	2 (.222)	9
PsycINFO	3 (.333)	1 (.111)	2 (.222)	3 (.333)	0 (.000)	9
Proadverbials						
INSPEC	12 (.571)	6 (.286)	2 (.095)	0 (.000)	1 (.048)	21
PsycINFO	10 (.555)	4 (.222)	0 (.000)	2 (.111)	2 (.111)	18
Subject(s)						
INSPEC	0 (.000)	0 (.000)	0 (.000)	0 (.000)	0 (.000)	0
PsycINFO	32 (.333)	3 (.031)	32 (.333)	11 (.115)	18 (.188)	96
Definite Article						
INSPEC	117 (.513)	53 (.232)	21 (.092)	16 (.070)	21 (.092)	228
PsycINFO	64 (.381)	28 (.167)	30 (.179)	26 (.155)	20 (.119)	168
TOTAL						
INSPEC	288 (.453)	147 (.231)	75 (.118)	49 (.077)	77 (.121)	636
PsycINFO	312 (.372)	129 (.154)	180 (.215)	111 (.132)	107 (.128)	839

When all subject experts' judgments are considered, the hypothesis is fairly well supported. Within the PsycINFO sample 37.2% of all anaphoric phrases are considered to be highly integral, and the percentage increases to 52.6% when the second category is included. Only 12.8% are judged to be peripheral, 26.0% if the adjacent category is included. The support is even stronger in the INSPEC sample, with 45.3% of anaphoric references judged to be integral, 68.4%, if the next category is included, while 12.1% are peripheral, 19.8% if the next category is included. Comparing highly integral with highly peripheral, there is roughly a 3 to 1 ratio in PsycINFO and a 4 to 1 ratio in INSPEC.

When the PsycINFO outlier's judgments are excluded, the data does not support the hypothesis so strongly. Those judged to be integral (in the first two categories) account for 41.8% of the data, and those judged to be peripheral (in the last two categories) increase to 32.6%. However, "Ss" ("subjects" in PsycINFO) comprise a sizable portion of the data (11.4%), and there is a noticeable shift in this category towards the peripheral side of the scale when the outlier is removed. Indeed, the outlier judged all occurrences of phrases replaced by "Ss" to be central to the abstract, while the others tended more toward peripheral. "Ss" is the most specific anaphor which we have encountered. By definition it represents a recognizable category of possible entities, whereas "this" or "which," for example, do not. It must refer to one or more physical beings which are or were under investigation. It seems that the importance of Subjects, which are often discussed in psychological literature, may be perceived quite differently by different researchers. If the class of "Ss" is removed from the analysis along with the outlier, the percentages improve to 45.1% for integral concepts and 31.6% for the peripherals.

No clear outlier among the INSPEC judges emerged. To test the worst case for supporting the hypothesis of the study, the subject expert who best supported the hypothesis, that is, the one with the highest proportion of "integral" judgments, was removed and the data were reanalyzed to see the effect on the centrality measure. Even then, the hypothesis is still well supported. Those anaphoric resolutions judged to be either highly central or nearly so still constitute 64.9% of the data while 23.5% were judged to be peripheral or nearly so.

Analyzing the results at the class level, it appears that six of the 10 classes of anaphors tend to replace integral concepts in abstracts. Approximately three fourths of the nominal demonstratives and proadverbials were judged central, although it is difficult to generalize from the small number of proadverbials in the sample. The definite article "the" is also judged integral in nearly three fourths of the INSPEC occurrences, and one half of the PsycINFO occurrences. Central pronouns and nominal substitutes are judged integral in two thirds of the INSPEC sample and in over half of the PsycINFO sample. The indefinites are also found to be integral in the PsycINFO sample. On the other hand, classes which are found to replace less integral concepts include the relative pronouns, proverbials, adjectives, and "Ss." It appears from the data analyzed in this study that although several classes of anaphors tend to refer to integral concepts in a body of text, other classes do not. Therefore, it may prove useful to resolve only certain anaphoric classes in order to better represent the "aboutness" of a document.

The differences noted between the INSPEC abstracts and the PsycINFO abstracts, when taking all classes into account, may well be due to the more frequent occurrence of relative pronouns and "Ss" in PsycINFO. These two classes, neither of which tend to refer to integral concepts as much as other classes, account for nearly one third of the occurrences of anaphors in PsycINFO as compared to only 14.2% of the INSPEC sample. This may

indicate a need to consider the sublanguage environment of the documents. The fact that there are fairly substantial differences between the two samples of documents which were analyzed for this study may be due to the size of the sample, but it may also indicate differences, such as the tendency to use certain classes of anaphors, which hold for entire collections of documents. Several studies [e.g., KITTREDGE82; TAGLIACOZZO76; TAGLIACOZZO78; SLOCUM86; BONZI in press] have demonstrated a variety of linguistic differences among various sublanguages. The use of anaphoric reference may be one such difference.

III. Effect on Term Weighting Schemes

The results of the test of the centrality of concepts referenced by anaphors suggest that the resolution of some classes of anaphors may be a potentially useful way in which to improve retrieval of documents in systems utilizing term weighting schemes. However, even if anaphors are to be resolved, it is necessary to know which weighting schemes most appropriately reflect the changes in frequencies. Some weighting schemes may not be affected to a great enough degree, while others are significantly affected, and still others may be negatively affected by the introduction of anaphoric resolutions. The most obvious case where a term weighting formula would be negatively affected is when the resolution of an anaphor introduces so many additional words that weights calculated with document length as a normalizing factor actually decrease when the resolved version of the document is used to calculate the weights of terms of interest.

This study tests the impact of anaphoric resolution on a variety of term weighting schemes only and does not deal at all with similarity measures. The question is merely how term weights are affected by anaphoric resolution. In an operational system, term weights must be considered in the context of the similarity between a user's query and the documents which are potentially relevant to the query. However, since similarity measures require term weights from queries as well as documents and we are investigating only the "aboutness" of documents and the effect of anaphors on term weights, we will not consider similarity measures. The data will serve to indicate which term weighting schemes reflect the changes when anaphors are resolved and which are not appropriate when anaphoric resolution is used. If a term's weight remains the same or even decreases after resolution of anaphors, then, regardless of the similarity measure used, the usefulness of anaphoric resolution is negligible.

Using 4 term weighting schemes, weights were calculated for each word which was part of an anaphoric resolution, first in the unresolved version of the text, and then in the resolved version. (Terms in the documents which were not part of the anaphoric resolutions were excluded because their weights would not change at all in two of the four term weighting schemes and minimally in the other two, unless they happened to occur in resolution.) Table 2 shows the mean weight for terms of interest in the unresolved and resolved abstracts and the average increase for each term weighting scheme.

As is evident from Table 2, the two term weighting schemes which do not incorporate document length (1 and 3) show a much greater proportional increase in term weights than do the two schemes which normalize on document length (2 and 4). In addition, the two schemes which use document length in their calculations produce a sizable number of virtually unchanged weights (calculated to the nearest hundredth) and even a few negative weights. Thus, it appears that if anaphoric resolution is used to better reflect the "aboutness" of a document, the term weighting schemes

used to measure that "aboutness" must be carefully chosen. This study indicates that term weighting schemes which take document length into account produce much different results from those which do not.

Table 2

Increase in Term Weights
with Anaphoric Resolution

	Unresolved	Resolved	Increase	Note
INSPEC (n=985)				
1. f	2.156	5.468	154%	
2. f/k	.024	.038	58%	18 neg. 199 unchanged
3. [f][log(N/d)]	8.114	19.440	140%	
4. $\frac{[f][\log(N/d)]}{k}$	.090	.139	54%	28 neg. 51 unchanged
PsycINFO (n=965)				
1. f	2.225	5.400	143%	
2. f/k	.021	.036	71%	10 neg. 335 unchanged
3. [f][log(N/d)]	8.348	19.835	138%	
4. $\frac{[f][\log(N/d)]}{k}$	.073	.133	82%	16 neg. 74 unchanged

f within document frequency
k tokens within the document
N documents in database
d postings

Performing an analysis of the results at the class level (but including the resolution of all anaphoric classes), similar proportions of increase in term weights hold, with a few notable exceptions. In the INSPEC sample, nominal substitutes show a much larger increase in weights in the two schemes which do not take document weights into account (253% and 240% increase), the same is true of the proadverbials (236% and 199% increase), and in PsycINFO, "subjects" shows a much larger increase in all term weights (324% and 308% increase in schemes not incorporating length, and 180% and 177% increase in those which do.)

IV. Effect of Resolving Individual Classes

The data were also analyzed by isolating the resolutions of five individual classes of anaphors in each sample. The classes were chosen first, because, within each discipline, they showed the highest degree of centrality as judged by the subject experts,and second, because there are enough cases to analyze. While the general analysis reported above was based on the resolution of all anaphors within the document, the following

is based on calculations using only the resolutions for the class of anaphors being analyzed. Table 3 shows the average increase in term weights for five anaphoric classes within each sample of data.

Table 3

Impact of Resolving Individual Classes of Anaphors

INSPEC

	Central Pronouns (n= 95)	Nominal Demonst. (n=178)	Nominal Substit. (n= 50)	Relative Pronouns (n= 97)	Definite Article (n=236)
1. f	62%	61%	84%	55%	77%
2. f/k	41% (11)	38% (33)	25% (23)	48% (13)	40% (53)
3. [f][log(N/d)]	62%	63%	79%	55%	80%
4. $\frac{[f][\log(N/d)]}{k}$	46% (5)	34% (14)	35% (5)	45% (2)	45% (16)

PsycINFO

	Central Pronouns (n= 97)	Nominal Demonst. (n=130)	Nominal Substit. (n= 43)	Indefinite Pronouns (n= 70)	Definite Article (n=131)
1. f	85%	68%	76%	53%	71%
2. f/k	68% (14)	25% (87)	25% (29)	29% (37)	38% (52)
3. [f][log(N/d)]	89%	69%	76%	52%	73%
4. $\frac{[f][\log(N/d)]}{k}$	76% (1)	43% (13)	45% (5)	41% (1)	49% (4)

f within document frequency
k tokens within the document
N documents in database
d postings

note: numbers in parentheses show number of words whose weight was either .00 or less, when the resolved text was measured.

In general, these individual classes exhibit smaller increases when only within class resolutions are included in the calculations. Obviously, this is because the increments in term frequency due to other resolutions within the document are not included. The impact of document length is also lessened, due to the smaller number of resolutions added in. Thus, the differences among the increases in term weights, as measured by the four term weighting schemes, are not as great as when the resolution of all classes is considered. It appears then, that if only selected anaphoric classes are resolved, there may be a wider choice of appropriate term weighting schemes which can be used to indicate the "aboutness" of the document since even those schemes which take document length into consideration are affected almost as strongly.

V. Effect of Resolving Combinations of Classes

The data were also manipulated to see how the term weighting schemes would react when two, three, four, or five classes of anaphors were resolved. Table 4 shows results when the classes are ordered according to the conceptual centrality of the resolution. For example, in the INSPEC sample, the definite article, followed by nominal demonstratives, tend to replace integral concepts to the greatest degree, so these two classes form the first group for analysis. Next are nominal substitutes, which are added to the first group to form the second group, and so on.

Table 4

Impact of Resolving Increasing Numbers of Classes of Anaphors According to Degree of Centrality

INSPEC

	Def. Art.+ Nom. Dem. (n=391)	Column 1+ Nom. Sub. (n=422)	Column 2+ Cen. Pro. (n=467)	Column 3+ Rel. Pro. (n=528)
1. f	75%	82%	89%	92%
2. f/k	38% (100)	38% (108)	42% (138)	43% (147
3. [f][log(N/d)]	78%	84%	89%	92%
4. $\frac{[f][\log(N/d)]}{k}$	37% (45)	34% (60)	37% (68)	38% (68)

PsycINFO

	Nom. Dem.+ Indef. (n=188)	Column 1+ Nom. Sub. (n=231)	Column 2+ Cen. Pro. (n=299)	Column 3+ Def. Art. (n=393)
1. f	69%	70%	86%	92%
2. f/k	30% (111)	32% (133)	53% (147)	47% (199)
3. [f][log(N/d)]	70%	70%	88%	94%
4. $\frac{[f][\log(N/d)]}{k}$	44% (13)	37% (37)	53% (43)	57% (44)

f within document frequency
k tokens within the document
N documents in database
d postings

note: numbers in parentheses show number of words whose weight was either .00 or less, when the resolved text was measured.

As might be expected, the addition of anaphoric classes increases the term weights with the first and third weighting schemes more than the second and fourth weighting schemes. The increases using the first and third schemes also appear to be more predictable. As each class is added, there is an increase in average weight. This is not necessarily the case with the second and fourth schemes. The same analysis was performed by selecting anaphoric classes which were most influenced by the term weighting schemes, and the same result was generally found. Overall, though, the increases were smaller, since the first two classes combined were those which had already shown to be the most influenced by term weighting schemes.

VI. Correlation Between Centrality and Term Weights

There appears to be good support for the hypothesis that term weighting schemes can be substantially influenced by anaphoric resolution. However, it is obvious that anaphoric resolution cannot be implemented indiscriminately, even with the use of appropriate term weighting schemes. There is not a strong correlation between a term's centrality to the document and its increase in term weight as a result of anaphoric resolution. Using Spearman's rho, correlations between subject experts' judgments of centrality of the phrases replaced by anaphors and a composite score of the proportional increase in term weights for all terms within the anaphoric resolution were calculated. In a class-by-class comparison with each of the four term weighting schemes investigated, the correlations range from strongly positive to strongly negative. In general, though, the correlations tend to be weakly positive. No term weighting scheme appears to correlate better or worse with the centrality score of the judges. Rather, the correlations tend to be tied to the anaphoric class. In PsycINFO, for example, there is a weakly positive correlation between the centrality of terms in an anaphoric resolution of central pronouns and the increase in term weights, while there is a somewhat stronger negative correlation between the centrality of nominal demonstratives and the increase in term weights. In the INSPEC sample, most correlations are positive, but these correlations still tend to be only weakly positive.

Although we cannot expect a general implementation of anaphoric resolution to be helpful in information retrieval, it appears that the correlation between centrality and increase in term weights is usually positive, and generally so with terms which tend to replace integral concepts.

VII. Conclusions

This study was deliberately removed from the "real world" of queries, documents retrieved in response to the queries, and relevance judgments of the retrieved documents. A user's relevance judgments are based on a number of factors, only one of which is the "aboutness" of a document. It is quite possible that a highly sophisticated system, one which resolves only anaphors which refer to integral concepts and ranks the documents with similarity measures using the best term weighting scheme to reflect the anaphoric resolution, will still fail its users. This can occur simply because "aboutness" is not the only dimension on which users base their relevance judgments. However, until we can devise systems which are able to discern and make use of these other dimensions, our best guess is still topicality.

We have become convinced, both from this and our previous investigation [KATZER86a], that anaphoric resolution does have applications in information retrieval. Results from this investigation indicate that anaphors do, in fact, tend to refer to integral concepts. Therefore, the

use of anaphoric resolution can give a better representation of the "aboutness" of a document, but the classes of anaphors to be resolved must be selectively chosen.

The varying impact of anaphoric resolution on a variety of term weighting schemes leads us to the conclusion that if anaphoric resolution is introduced into an information retrieval system, care must be taken in choosing a term weighting scheme which reflects the increase in the frequency of terms of potential interest. The study has shown that schemes which do not use document length as part of the calculations tend to show the increases in the terms affected by anaphoric resolution much better than those which do. However, when only individual classes of anaphors are resolved, the differences among the term weighting schemes become smaller.

This work was supported by Syracuse University Senate Committee on Research Award No. 40.

References

[BONZI88] Bonzi, S. "Syntactic Patterns in Scientific Sublanguages: A Study of Four Disciplines." Journal of the American Society for Information Science. (In press)

[KATZER86a] Katzer, J., S. Bonzi, and E. Liddy. Impact of Anaphoric Resolution in Information Retrieval. Syracuse, New York: Syracuse University School of Information Studies. Final Report to the National Science Foundation.

[KATZER86b] Katzer, J., S. Bonzi, and E. Liddy. "The Effects of Anaphoric Resolution on Retrieval Performance: Preliminary Findings." ASIS '86: Proceedings of the 49th ASIS Annual Meeting. Volume 23. Medford, N.J.: Learned Information, Inc.

[KITTREDGE82] Kittredge, R. "Variation and Homogeneity of Sublanguages." In: R. Kittredge and J Lehrberger, ed. Sublanguage: Studies of Language in Restricted Semantic Domains. New York: Walter de Gruyter, pp. 107-37.

[LIDDY87] Liddy, E., S. Bonzi, J. Katzer, and E. Oddy. "A Study of Discourse Anaphora in Scientific Abstracts." Journal of the American Society for Information Science. 38(4):255-61.

[O'CONNOR73] O'Connor, J. "Text Searching Retrieval of Answer-Sentences and other Answer-Passages." Journal of the American Society for Information Science. 24(6):445-60.

[PAICE81] Paice, C.D. "The Automatic Generation of Literature Abstracts: An Approach Based on the Identification of Self-Indicating Phrases." In: R.N. Oddy, ed. Information Retrieval Research. London: Butterworths. pp. 172-91.

[PAICE87] _____, and G.D. Husk. "Toward the Automatic Recognition of Anaphoric Features in English Text: The Impersonal Pronoun 'it'." Lancaster, U.K.: University of Lancaster Department of Computing.

[SALTON83] Salton, G. and M.J. McGill. Introduction to Modern Information Retrieval. New York: McGraw-Hill.

[SLOCUM86] Slocum, J. "How One Might Automatically Identify and Adapt to a Sublanguage: An Initial Exploration." In: R, Grishman and R. Kittredge, eds. Analyzing Language in Restricted Domains: Sublanguage Description and Processing. Hillsdale, N.J.: Lawrence Erlbaum Associates.

[TAGLIACOZZO76] Tagliacozzo, R. "Levels of Technicality in Scientific Communication." Information Processing and Management. 12:95-110.

[TAGLIACOZZO76] _____. "Some Stylistic Variations in Scientific Writing." Journal of the American Society for Information Science. 29:136-40.

Overcrowding in the home: An empirical investigation of possible pathological consequences.

Several recent studies have suggested that, contrary to investigators' initial expectations, household crowding typically has little impact on humans. Using a sample (a total of 2,035 Ss interviewed) collected in Chicago minimized the collinearity between crowding and socioeconomic variables, the authors found that both objective crowding (as measured by persons per room) and subjective crowding (as indicated by excessive social demands and a lack of privacy) were strongly related to poor mental health, poor social relationships in the home, and poor child care; and were less strongly but significantly related to poor physical health and poor social relationships outside the home. 3 crowding variables taken together, on the average, uniquely explain as much variance in this study's dependent variables as is uniquely explained the combined effects of sex, race, education, income, age, and marital status. It is suggested that attention be turned away from the question of whether crowding ever has effects to the study of factors maximize or minimize its effects.

1 2 3 4 5	Overcrowding in the home
1 2 3 4 5	a sample (a total of 2,035 Ss interviewed) collected in Chicago
1 2 3 4 5	persons per room
1 2 3 4 5	excessive social demands and a lack of privacy
1 2 3 4 5	poor mental health, poor social relationships in the home, and poor child care
1 2 3 4 5	poor physical health and poor social relationships outside the home
1 2 3 4 5	factors