THE EFFECTIVENESS OF THE THESAURUS METHOD IN AUTOMATIC INFORMATION RETRIEVAL

C.T. Yu[+] and G. Salton*

TR 75-261

November 1975

Department of Computer Science
Cornell University
Ithaca, New York

[+]Department of Computing Science, University of Alberta, Edmonton, Alberta.

*Department of Computer Science, Cornell University, Ithaca, NY 14853.

This study was supported in part by the Canadian Research Council and in part by the National Science Foundation under grant GJ 43505.

The Effectiveness of the Thesaurus Method
in Automatic Information Retrieval

C.T. Yu[+] and G. Salton[*]

Abstract

Term grouping and thesaurus methods have frequently been incorporated into automatic content analysis programs as devices for the recognition of synonymous expressions and of linguistic entities that may be semantically similar but syntactically distinct. While it has frequently been asserted that the recognition of synonyms is essential in language analysis, actual proofs of the usefulness of a thesaurus in automatic information retrieval are outstanding.

In the present study, formal proofs are given of the effectiveness under well-defined conditions of the thesaurus method in information retrieval. It is shown, in particular, that when certain semantically related terms are added to the information queries originally submitted by the user population, a superior retrieval system is obtained in the sense that for every level of the recall the retrieval precision is at least as good for the altered queries as for the original ones.

1. Introduction

A good deal is known about the representation of document content and the assignment of effective content identifiers (index terms, keywords, descriptors)

[+] Department of Computing Science, University of Alberta, Edmonton, Alberta.

[*] Department of Computer Science, Cornell University, Ithaca, NY 14853.

This study was supported in part by the Canadian Research Council and in part by the National Science Foundation under grant GJ 43505

to documents and information requests. Among the characteristics of good content identifiers the following are now widely agreed upon by experts in the field:

a) Good content bearing words tend to occur in the documents of a collection with uneven frequency distributions; that is, in certain documents their occurrence frequencies are much larger than would be expected from a random assignment of terms to documents; nonspecialty words, on the other hand exhibit random occurrence patterns in the documents of a collection.

b) The most effective content identifiers exhibit little redundancy with other terms also used for content identification; in particular, terms with high document frequency — those assigned to a large proportion of the documents of a collection — tend to be indiscriminate in their retrieval capability and lead to losses in retrieval precision.*

c) Effective content identifiers are expected to break up large clusters of documents that are not otherwise distinguishable for retrieval purposes; that is, they should reduce the existing uncertainty for the given document set. Thus, terms that occur with excessively low document frequency in the documents of a collection are not optimal and lead to unacceptable losses in recall.

---

* The effectiveness of a retrieval system is often evaluated by two complementary measures known as precision and recall, respectively, defined as the proportion of retrieved items that are relevant, and the proportion of relevant items that are retrieved. In general, an effective retrieval system exhibits high values for both recall and precision in that the user expects to retrieve a reasonable proportion of what is relevant while at the same time rejecting a high proportion of what is extraneous.

These considerations have given rise to a variety of automatic indexing strategies designed to assign appropriate content identifiers to the documents of a collection and to incoming user queries. One such is the <u>discrimination value method</u> which has been used with a variety of document collections in different subject areas. [1,2] The best discriminators are invariably found to be terms of average document frequency — they occur normally in more than one one-hundredth of the documents of a collection, but less than one tenth of the collection. In the discrimination value model the good discriminators are assigned as content identifiers to documents and queries without any modifying transformation.

Terms whose document frequency is either too high or too low often lead to unacceptable losses in precision and recall, respectively, and must be transformed into better terms by an appropriate reduction (or increase) in their document frequencies. Two types of frequency transformations are therefore introduced:

a) a decreasing frequency transformation applicable to the high frequency terms which by combining such terms into term <u>phrases</u> produces content identifiers of lower document frequency that are more specific than the original phrase components;

b) an increasing frequency transformation applicable to the low frequency terms which assembles such terms into <u>classes</u> of similar or related terms; by assigning such term or <u>thesaurus classes</u> as content identifiers, higher frequency, more general entities are produced than the original class entries.

The main role assigned to the thesaurus by the discrimination value model is then as a device for assembling low frequency terms into classes in the hope of creating more general content identifiers that lead to improvements in the recall performance.

2. The Thesaurus Method

Before embarking on the mathematical development, it may be useful briefly to outline the proof procedures and the assumptions leading to the results.

Query and document vectors are assumed to be binary, that is, d_i [q_i] equals 1 whenever term i is present in document D [query Q], and is zero otherwise. The similarity function s between queries and documents is assumed to be

$$s(D, Q) = \sum_{i=1}^{n} d_i\, q_i$$

where n is the vector length (the number of distinct terms in the vectors). For binary vectors, s represents the number of matching terms between the query and document vectors, respectively.

The evaluation of the effectiveness of a particular method of term assignment is based on the comparison of the retrieval precision at given levels of the recall. Consider a specified recall level γ (a specified proportion of relevant items retrieved), and let $|R|$ be the total number of relevant items for a given query. Then the precision P_γ at recall level γ may be defined as

$$P_\gamma = \frac{\gamma|R|}{\text{Total number of items to be retrieved in order to obtain } \gamma|R| \text{ relevant ones}} .$$

A retrieval system (A) is then assumed to be <u>superior</u> to an alternative system (B) if and only if for all recall levels γ, the retrieval precision for (A) is at least as large as that for (B).

The computation of P_Y makes it necessary to identify the number of nonrelevant documents that must be retrieved for each increase of 1 in the number of relevant documents obtained. This in turn requires the following assumptions to be made regarding the occurrences of terms in the documents of the collection and the composition of the relevant and nonrelevant document sets for each query:

Assumption 1: For each query, the corresponding query terms are assumed to be independently assigned to the documents of the collection. Furthermore the terms are assumed to be uniformly distributed across the set of relevant documents R and the set of nonrelevant documents I. That is, the probability of occurrence of a given term j_k has the same value for all relevant documents in R; similarly the value is the same for all nonrelevant documents in I (although the two probabilities may differ among themselves).

Thus, if one assumes that the probability of a relevant [nonrelevant] document containing term j_k is $r_{jk}/|R|$ $[\sigma_{jk}/|I|]$, where r_{jk} and σ_{jk} are the number of relevant and nonrelevant documents, respectively, containing term j_k, then the probability that a given relevant [nonrelevant] contains a given term set $(j_1, j_2, \ldots, j_p)$ will be $\prod_{k=1}^{p} r_{jk}/|R|$ $[\prod_{k=1}^{p} \sigma_{jk}/|I|]$.

Assumption 2: All documents exhibiting a given number of matching query-document terms have equal chance of being retrieved. That is, if c $(c \geq 1)$ relevant items and g nonrelevant items all exhibit the same similarity coefficient with respect to some query Q, then it is assumed that g/c nonrelevant items are retrieved for each relevant retrieved. That is, the relevant items occur at even intervals among the nonrelevant in the ranked list of retrieved documents (the ranking is assumed in decreasing order of the query-document similarity).

The intention of the thesaurus method is to create from each original query Q a new query Q^* obtained by adding to Q one or more terms that are "semantically related" to the original terms. More specifically, for each term q included in Q, a set of related terms is defined as the set of all terms included in the same thesaurus class as q. All such related terms are then added to Q to form Q^*. Each of the new terms $\{j_1, j_2, \ldots, j_\ell\}$ added to the query $Q = \{1, 2, \ldots, m\}$ is weighted by a factor Δ/ℓ, where $\Delta < 1$. This means that the increment in the similarity between Q^* and D due to the added terms will be strictly less than 1.

One additional restriction applies to the terms supplied by the thesaurus, motivated by its role as a classification of low frequency, specific terms. The thesaurus terms must be "high precision" terms, that is, their probability of occurrence in the documents relevant to a given query must not be smaller than their probability of occurrence in the nonrelevant items. More precisely, for each term j_k included in the thesaurus and for each query Q

$$r_{j_k}/|R| \geq \sigma_{j_k}/|I|.$$

There is considerable evidence that "term precision" as defined here is inversely related to document frequency, and that for the low-frequency terms included in a thesaurus, this requirement is satisfied in most cases. [3]

The main theorem may now be stated as follows: the thesaurus method providing for the addition to the original user queries of semantically related terms taken from a thesaurus produces a superior retrieval system.

The relevant proof appears in the next section.

3. Thesaurus Effectiveness

The main proof makes use of a technical lemma which may be stated as follows: consider a function of ℓ terms $S_1, S_2, \ldots, S_\ell$ ($0 \leq S_i \leq 1$, $i = 1, 2, \ldots, \ell$) consisting of sums of products of ℓ terms each, each product containing j factors ($j \leq \ell$) chosen from the set of S_i and ℓ-j terms ($\ell - j \geq 0$) consisting of factors $(1 - S_i)$. Specifically, let

$$c(S_1, S_2, \ldots, S_\ell; j) = \Sigma \left[\prod_{k=1}^{j} S_\rho(k) \right] \left[\prod_{k=j+1}^{\ell} (1-S_\rho(k)) \right]$$

where ρ denotes a permutation of $\{1, 2, \ldots, \ell\}$ and the summation covers all the $\binom{\ell}{j}$ combinations of j terms out of ℓ.*

Then, if $S_g \geq S_g'$, one has

$$\sum_{j=t}^{\ell} c(S_1, S_2, \ldots S_{g-1}, S_g, S_{g+1}, \ldots S_\ell; j)$$

$$\geq \sum_{j=t}^{\ell} c(S_1, S_2, \ldots S_{g-1}, S_g', S_{g+1}, \ldots S_\ell; j).$$

The proof appears in the appendix.

---

* For example $c(S_1, S_2, S_3, S_4; 3) = S_1 S_2 S_3 (1 - S_4) + S_1 S_2 S_4 (1 - S_3) + S_1 S_3 S_4 (1 - S_2) + S_2 S_3 S_4 (1 - S_1)$.

The main theorem is true if for every recall point γ the thesaurus method provides a retrieval precision which is not inferior to that obtainable with the standard (nonthesaurus) process. In a retrieval situation in which the retrieved documents are presented to the user in decreasing order of the corresponding query-document similarity coefficient, a recall, and hence a precision, value may be calculated following the retrieval of each individual document (that is, after retrieval of the first item; after the second item; after the third item, and so on, down to the last retrieved document).

Among all the recall values obtained in this way, some are of special interest corresponding to the retrieval of the last document within each set of documents exhibiting a common number of matching query-document terms and including a relevant item; these special recall levels are known as <u>standard recall points</u>. A typical example showing a ranked list of retrieved documents is shown together with its standard recall points in Table 1.

The main theorem will be proved first for the standard recall points, and later for any nonstandard recall level situated between adjacent standard recall points.

<u>Theorem</u>: The thesaurus method provides a superior retrieval system.

<u>Proof</u>: Consider the situation first for all standard recall points q_i for which all documents with query document similarity greater or equal to i are retrieved by the original query Q.

Any document D_j not retrieved by Q at recall point q_i has similarity equal at most to $(i-1) + \Delta$ with Q^*, where $(i-1) + \Delta < i$. Such a document is then not retrieved by Q^* at standard recall point q_i. On the other hand any relevant document D_k retrieved by Q at q_i will necessarily exhibit a similarity coefficient with Q^* at least equal to i. Thus all relevant documents retrieved by Q at q_i are also retrievable by Q^*; at the same time nonrelevant items not retrieved by Q at q_i are also rejected by Q^*.

Consider now an arbitrary nonstandard recall point x situated between q_i and the preceding standard recall point q_{i+1}. The documents retrieved at recall point x fall into two classes

i) those whose similarity with Q is at least equal to i+1,

and ii) those whose similarity with Q is exactly equal to i.

Let B' and B" be the number of relevant and nonrelevant documents of type (i) respectively. Analogously, let X' and X" be the number relevant and nonrelevant items of type (ii). If p, $0 < p \leq X'$, is the number of relevant documents of type (ii) retrieved by Q at recall point x, then the total number of retrieved documents (both relevant and nonrelevant) of type (ii) will be $(p/X') \cdot (X'+X'')$ since by Assumption 2 all documents with a given number of query-document term matches are assumed to be retrievable equally easily. The precision for Q at recall point x is then

$$\frac{B' + p}{B' + B'' + \frac{p}{X'}(X' + X'')} \quad . \qquad (1)$$

Two types of documents also exist for query Q^* at recall point x, namely

iii) those whose similarity with Q^* is i + 1 or larger,

and iv) those whose similarity with Q^* is at least equal to i but less than i + 1.

For Q^*, however, the documents of type (iv) are further subdivided into $\ell+1$ subclasses, including those whose similarity coefficient with Q^* equals $i + \Delta$, $i + (\frac{\ell - 1}{\ell})\,\Delta$, ..., $i + \frac{\Delta}{\ell}$, i. Let the number of relevant (nonrelevant) documents in the $\ell + 1$ different subclasses be X_1' (X_1''), X_2' (X_2''), ..., $X_{\ell+1}'$ ($X_{\ell+1}''$), respectively, with X_1' corresponding to similarity coefficient $i + \Delta$, and $X_{\ell+1}$ to similarity i.

Obviously, the X_1' relevant documents exhibit i matches with terms originally included in Q and ℓ matches with the added terms $\{j_1, j_2, \ldots, j_\ell\}$. The same is true for the X_1'' nonrelevant documents. The remaining document classes exhibit correspondingly fewer matches with the added terms.

Since by Assumption 1 the distribution of query terms is assumed uniform across all relevant documents, and terms are independently assigned, it is clear that

$$X_1' = X' \cdot \left(\prod_{k=1}^{\ell} \frac{r_{jk}}{|R|}\right)$$

$$= X' \cdot c\left(\frac{r_{j1}}{|R|}, \frac{r_{j2}}{|R|}, \ldots, \frac{r_{j\ell}}{|R|}; \ell\right),$$

where $r_{jk}/|R|$ is the probability that a relevant document contains term j_k. Similarly, X_2', ..., $X_{\ell+1}'$ will be equal respectively to $X' \cdot c\left(\frac{r_{j1}}{|R|}, \ldots, \frac{r_{j\ell}}{|R|}; \ell-1\right)$, ..., $X' \cdot c\left(\frac{r_{j1}}{|R|}, \ldots, \frac{r_{j\ell}}{|R|}; 0\right)$.

Without loss of generality consider p, the number of relevant documents retrieved by Q at recall point x with i matching terms, such that $\sum_{k=1}^{\mu+1} X_k' > p \geq \sum_{k=1}^{\mu} X_k'$ for some integer μ, $0 < \mu \leq \ell + 1$. The proof is given first for $p = \sum_{k=1}^{\mu} X_k'$. For such a value of p, the precision value for Q^* will be

$$\frac{B' + p}{B' + B'' + \sum_{k=1}^{\mu} (X_k' + X_k'')}. \qquad (2)$$

By comparing the denominators of (1) and (2), the result follows provided that

$$\frac{p}{X'}(X' + X'') \geq \sum_{k=1}^{\mu}(X_k' + X_k'').$$

Since $p = \sum_{k=1}^{\mu} X_k'$, this implies

$$\frac{X' + X''}{X'} \geq \frac{\sum_{k=1}^{\mu}(X_k' + X_k'')}{\sum_{k=1}^{\mu} X_k'}$$

or again

$$\frac{X'}{X''} \leq \frac{\sum_{k=1}^{\mu} X_k'}{\sum_{k=1}^{\mu} X_k''} \quad . \tag{3}$$

The summations can be replaced as shown earlier as follows

$$\frac{X'}{X''} \leq \frac{X'}{X''} \cdot \left\{ \frac{\sum_{k=\ell-\mu+1}^{\ell} c\left(\frac{r_{j1}}{|R|}, \ldots, \frac{r_{j\ell}}{|R|}; k\right)}{\sum_{k=\ell-\mu+1}^{\ell} c\left(\frac{\sigma_{j1}}{|I|}, \ldots, \frac{\sigma_{j\ell}}{|I|}; k\right)} \right\} \tag{4}$$

Expression (4) is obviously true provided the sum in the numerator exceeds that of the denominator, that is, provided

$$\sum_{k=\ell-\mu+1}^{\ell} c\left(\frac{r_{j1}}{|R|}, \ldots, \frac{r_{j\ell}}{|R|}; k\right) \geq \sum_{k=\ell-\mu+1}^{\ell} c\left(\frac{\sigma_{j1}}{|I|}, \ldots, \frac{\sigma_{j\ell}}{|I|}; k\right).$$

But, by the term precision assumption $\frac{r_{jk}}{|R|} \geq \frac{\sigma_{jk}}{|I|}$, $1 \leq k \leq \ell$.
Thus a repeated application of the results of the lemma establishes the result.

Consider now $\sum_{k=1}^{\mu+1} X_k' > p > \sum_{k=1}^{\mu} X_k'$. The precision of the augmented query Q^* at recall point x equal to $(B' + p)/|R|$ will be

$$\frac{B' + p}{B' + B'' + \sum_{k=1}^{\mu} (X_k' + X_k'') + \left[\frac{p - \sum_{k=1}^{\mu} X_k'}{X_{\mu+1}'} \right] (X_{\mu+1}' + X_{\mu+1}'')} \quad (5)$$

The denominator of (5) includes all retrieved documents exhibiting at least $i + 1$ original term matches with Q^* (that is, $B' + B''$), followed by the documents with i original term matches and up to $\ell - \mu + 1$ matches through the added terms $\{j_1, j_2, \ldots, j_\ell\}$. The right-most term in the denominator of (5) covers a subset of the documents exhibiting i original term matches and $\ell - \mu$ matches through the added terms.

By comparing (1) and (5), it is seen that the performance of Q^* at recall point x will be at least as good as that of the original query Q if and only if

$$\frac{p}{X'}(X' + X'') \geq \sum_{k=1}^{\mu} (X_k' + X_k'') + \frac{p - \sum_{k=1}^{\mu} X_k'}{X_{\mu+1}'} (X_{\mu+1}' + X_{\mu+1}'').$$

This is equivalent to

$$p\left(\frac{X''}{X'} - \frac{X_{\mu+1}''}{X_{\mu+1}'}\right) - \sum_{k=1}^{\mu} X_k'' + \sum_{k=1}^{\mu} X_k' \left(\frac{X_{\mu+1}''}{X_{\mu+1}'}\right) \geq 0.$$

By adding $\sum_{k=1}^{\mu} X_k' \cdot \frac{X''}{X'} - \sum_{k=1}^{\mu} X_k' \cdot \frac{X''}{X'}$ to the previous expression, one obtains

$$p\left\{\frac{X''}{X'} - \frac{X_{\mu+1}''}{X_{\mu+1}'}\right\} - \sum_{k=1}^{\mu} X_k' \left\{\frac{X''}{X'} - \frac{X_{\mu+1}''}{X_{\mu+1}'}\right\} + \sum_{k=1}^{\mu} \left(X_k' \cdot \frac{X''}{X'} - X_k''\right) \geq 0$$

or finally

$$\left(p - \sum_{k=1}^{\mu} X_k'\right)\left(\frac{X''}{X'} - \frac{X_{\mu+1}''}{X_{\mu+1}'}\right) + \sum_{k=1}^{\mu} \left(X_k' \cdot \frac{X''}{X'} - X_k''\right) \geq 0 \qquad (6)$$

Since $\frac{X'}{X''} \leq \frac{\sum_{k=1}^{\mu} X_k'}{\sum_{k=1}^{\mu} X_k''}$ by equation (3), the second term of (6) is obviously greater or equal to 0. The first factor of the left-hand term in (6) is greater than zero since $p > \sum_{k=1}^{\mu} X_k'$ for the case under consideration. Thus if $X''/X' \geq X_{\mu+1}''/X_{\mu+1}'$ the theorem is established. If on the other hand $X''/X' < X_{\mu+1}''/X_{\mu+1}'$, the first term of (6) becomes negative since the two factors of the product have opposite sign. By substituting in (6) a larger value of p than that for the current case (for example, $\sum_{k=1}^{\mu+1} X_k'$), a new expressi[on] is obtained which is necessarily smaller than (6):

$$\left(\sum_{k=1}^{\mu+1} X_k' - \sum_{k=1}^{\mu} X_k'\right)\left(\frac{X''}{X'} - \frac{X''_{\mu+1}}{X'_{\mu+1}}\right) + \sum_{k=1}^{\mu} \left(X_k' \frac{X''}{X'} - X_k''\right) \geq 0. \qquad (7)$$

But expression (7) covers the previously treated case where $p = \sum_{k=1}^{\mu} X_k'$ for some integer μ; for that case the theorem has already been proved. Thus (7) is reducible to (3) and the proof is complete.

The proof procedure given here for the thesaurus method is usable under somewhat different assumptions and conditions for other retrieval techniques including term weighting and phrase transformations. [3,4]

References

[1] G. Salton, C.S. Yang, and C.T. Yu, A Theory of Term Importance in Automatic Text Analysis, Journal of the ASIS, Vol. 26, No. 1, January - February 1975, p. 33-44.

[2] G. Salton, A Theory of Indexing, Regional Conference Series in Applied Mathematics No. 18, Society for Industrial and Applied Mathematics, Philadelphia, 1975.

[3] C.T. Yu and G. Salton, Precision Weighting — An Effective Automatic Indexing Method, to be published in Journal of the ACM; also Technical Report No. 75-232, Department of Computer Science, Cornell University, 1975.

[4] C.T. Yu and G. Salton, Effective Information Retrieval Using Term Accuracy, Technical Report No. 75-249, Department of Computer Science, Cornell University, 1975.

Document Rank	Relevance Indicator (R means relevant)	Number of Matching Terms	Standard Recall Point ✓	Recall	Precision
1	R	7		0.1	
2	N	7	✓	0.1	0.5
3	N	6		0.1	
4	N	6		0.1	
5	R	5		0.2	
6	N	5		0.2	0.33
7	R	5		0.3	
8	N	5	✓	0.3	0.37
9	R	4		0.4	
10	N	4		0.4	0.42
11	R	4	✓	0.5	0.45
12	N	3			
.	.	.	.	.	
.	.	.	.	.	
.	.	.	.	.	

Typical Precision Computation at

Standard Recall Points

(assumption: total number of relevant is 10)

Table 1

✓ standard recall points

Appendix

Lemma: If $S_g \geq S_g'$ it follows that

$$\sum_{j=t}^{\ell} C(S_1, S_2, \ldots, S_{g-1}, S_g, S_{g+1}, \ldots, S_\ell; j) \geq$$

$$\sum_{j=t}^{\ell} C(S_1, S_2, \ldots, S_{g-1}, S_g', S_{g+1}, \ldots, S_\ell; j).$$

Proof:

$$\begin{aligned}
& C(S_1, S_2, \ldots, S_{g-1}, S_g, S_{g+1}, \ldots, S_\ell; j) \\
= \quad & S_g \cdot C(S_1, S_2, \ldots, S_{g-1}, S_{g+1}, \ldots, S_\ell; j-1) \\
& + (1-S_g) \cdot C(S_1, S_2, \ldots, S_{g-1}, S_{g+1}, \ldots, S_\ell; j) \\
= \quad & S_g \big[C(S_1, S_2, \ldots, S_{g-1}, S_{g+1}, \ldots, S_\ell; j-1) \\
& - C(S_1, S_2, \ldots, S_{g-1}, S_{g+1}, \ldots, S_\ell; j)\big] \\
& + C(S_1, S_2, \ldots, S_{g-1}, S_{g+1}, \ldots, S_\ell; j).
\end{aligned}$$

Thus

$$\begin{aligned}
& \sum_{j=t}^{\ell} C(S_1, S_2, \ldots, S_{g-1}, S_g, S_{g+1}, S_\ell; j) \\
= \quad & S_g \sum_{j=t}^{\ell} \big[C(S_1, S_2, \ldots, S_{g-1}, S_{g+1}, \ldots, S_\ell; j-1) \\
& - C(S_1, S_2, \ldots, S_{g-1}, S_{g+1}, \ldots, S_\ell; j)\big] \\
& + \sum_{j=t}^{\ell} C(S_1, S_2, \ldots, S_{g-1}, S_{g+1}, \ldots, S_\ell; j).
\end{aligned}$$

The factor $C(S_1, S_2, \ldots, S_{g-1}, S_{g+1}, \ldots, S_\ell;\ \ell)$ can be defined as **zero**, because one cannot factor out ℓ terms when only $\ell - 1$ are present. Furthermore, all but the first term appearing in the square brackets cancel; that is

$$\sum_{j=t}^{\ell} [C(S_1, S_2, \ldots, S_{g-1}, S_{g+1}, \ldots, S_\ell;\ j-1) - C(S_1, S_2, \ldots, S_{g-1}, S_{g+1}, \ldots, S$$

$$= C(S_1, \ldots;\ t-1) - C(S_1, \cancel{\ldots;\ } t) + C(S_1, \cancel{\ldots;\ } t)$$

$$- C(S_1, \cancel{\ldots;\ } t+1) + \ldots\ldots - C(S_1, \cancel{\ldots,\ } \ell)]$$

$$= C(S_1, \ldots, S_\ell;\ t-1).$$

Thus

$$\sum_{j=t}^{\ell} C(S_1, S_2, \ldots, S_{g-1}, S_g, S_{g+1}, \ldots;\ S_\ell;\ j)$$

$$= S_g \cdot C(S_1, S_2, \ldots, S_{g-1}, S_{g+1}, \ldots, S_\ell;\ t-1)$$

$$+ \sum_{j=t}^{\ell-1} C(S_1, S_2, \ldots, S_{g-1}, S_{g+1}, \ldots, S_\ell;\ j). \qquad (4)$$

The lemma is an immediate consequence of the last expression. ∎