人文科学とコンピュータ 16－6
（１９９２．１１．２７）

# 機械可読テキストにおける文脈の基本演算について

小沢　一雅
（大阪電気通信大学工学部）

キーワードによって定義される文脈領域に対する基本演算として、ＡＮＤ、ＯＲおよびＣＯＲなる３つの演算を新たに導入する。各演算の厳密な定義および基本的な性質を考えるとともに、機械可読テキストからの情報検索への適用について理論的・実験的考察を行う。とくに、検索木による検索実験を通じて各演算の有効性を調べる。実験に用いるテキストとして「マルコによる福音書」をとりあげ、たとえば、「イエスはオリーブ山で何を言ったか？」のような問い合わせに対する検索実験を実施した。

# Contextual Operations for Information Retrieval from Machine Readable Texts

**Kazumasa Ozawa**
Department of Management Engineering
Osaka Electro-Communication University
Neyagawa-Shi, Osaka 572, Japan
(email: ozawa@ozlab.osakac.ac.jp)

*Abstract:* Newly defined Boolean-like AND, OR and COR operations between contextual domains are theoretically and experimentally discussed, which play key roles in information retrieval from machine readable texts. The contextual domain of a given keyword is defined in a non-gramatical way such as KWIC. Query operations are experimentally carried out on a small-sized text; The Gospel According to Mark. An example of query is "What did Jesus say on the Mount of Olives?". The results show that the proposed OR and COR operations act very well in full text retrieval.

*Keywords:* Machine Readable Text, Full Text, Information Retrieval, Context, Operation, KWIC, Database.

## 1. Introduction

Many machine readable texts have recently been published in a variety of fields, gradually changing our traditional style of research, investigation or document analysis. Even today, people still read books and printed documents by taking in their hands. Such a reading style, however, appears to be rapidly limited within a small domain as enjoying nobels, newspapers or letters from friends.

It is often seen that a researcher turns over the leaves of a bulky volume, like a busy bee, slipping many markers between leaves. In this situation, he does not *read*, but *retrieves* some information from the volume. If it was published as a machine readable text, there would be much better ways of information retrieval. One scheme would be brought by expanding the existing text processing techniques into our problem domain: Some fast string matching algorithms have been presented for text processing [1-3]. The KWIC (Keyword-in-Context) indexing has widely been used since it was first introduced in a special field [4]. These techniques have been playing key roles in full text retrieval or analysis [5-7].

This paper presents newly defined Boolean-like operations on a full text, which act well in the keyword-based information retrieval from the text. The operations are closely related to the contexts associated with given keywords. Contextual methods for text processing have been presented in many papers. Among them, one approach is grammatical; contexts are regarded as grammatical units like phrases [8]. Another is non-grammatical, as seen in KWIC indices; contexts are defined by strings composed by a number of symbols. In this paper, all discussions will be presented in terms of non-grammatical contexts.

Every machine readable text is not well-structured like a database. It follows that any full text retrieval is more or less incomplete. This paper also presents a discussion on loss of information caused by our query operations.

## 2. Contexts and information retrieval

We aim at building a new scheme to retrieve information from a machine readable text in a similar way as from a database. For example, suppose that we have a machine readable text of a bulky volume on the world history. If it was a structured database, we would be able to make inquiries about historical events such as:

"What did Caesar in 60 BC?"
"What happened when Jinghis Khan attacked Baghdad?"
"Who met the Emperor of Ch'in in 215 BC?"

In order to make it possible to ask such questions to the text, we have to detect correlation between keywords, for example, "Caesar" and "60 BC". Any word can not directly correlate with other words, but does indirectly through the medium of contexts. KWIC indices clearly show that any word can not fix its meaning without the context. No word stands independently of its context. Then it follows that correlation between words is nothing but one between their contexts. Consider again the first of the three questions presented above: Suggestive information to answer the question is likely to be detected around a part where two contexts associated with "Caesar" and "60 BC" are interfered with each other. To apply this into practice, our intuitive and ambiguous concepts like *text*, *context*, *correlation* and so on should be formally defined: Precise definitions will be presented in the next chapter.

## 3. Definitions

First we define a machine readable text, simply termed *text*, as a string of symbols. Let T be a text. We have

$$T = "a_1 a_2 \ldots . a_n".$$

Where $a_1, a_2, \ldots, a_n$ are the n symbols composing T. We call $a_1$ and $a_2$ the *start* and *terminal symbols*, respectively. The *length* of T, i.e. the number of symbols in T, is written by #T. The natural number i of symbol $a_i$ $(i=1, \ldots, n)$ is called a *position*. Here, a *section* is defined by a set of successive positions. Symbolically, we write

$$[i, j] = \{i, i+1, i+2, \ldots, j\} \quad (i \leq j).$$

Where the minimum position i and the maximum position j are called the *left* and *right ends* of the section, respectively. For T such that #T=n, the longest section U=[1, n], termed the *universe*, can be considered. Any section [i, j]

is included by U in terms of set theory. Symbolically, [i, j] $\subset$ U. A string of symbols placed on a section d=[i, j] is denoted by s[i, j] or s[d]. Symbolically, we denote

$$s[i, j] = "a_i a_{i+1} a_{i+2} \ldots . a_j".$$

**Definition 1** For any two sections d and d', we can define the following operations in terms of set operations. Where, for disjoint d and d', we have a peculiarity such that the operational result does not mean a section:

(a) For d and d' having position(s) in common, we have:

d OR d' (*Union*; producing a unified section)

d AND d' (*Intersection*; producing a section of common positions)

(b) For d and d' having no position in common (*disjoint*), we have:

d OR d' (Union; producing a set of sections {d, d'})

d AND d' (Intersection; producing the null section $\varnothing$)

(c) For any d and d', we have a new operation defined by:

$$d \text{ COR } d' = \begin{cases} d \text{ OR } d' & \text{for } d \text{ AND } d' \neq \varnothing \\ \varnothing & \text{for } d \text{ AND } d' = \varnothing \end{cases} \qquad (1)$$

The operation defined by (1), named *COR(Correlating OR)*, plays very important role in our scheme of information retrieval.

**Definition 2** For a section d=[i, j] in U=[1, n], an *extention operator* ext(p) is defined as follows:

$$ext(p)d = [i-p, j+p] \qquad (2)$$

Where, for i-p<1, the left end is replaced with 1 and, for j+p>n, the right end is replaced with n. For simplicity, the extended d, i.e. ext(p)d, is written by *d.

**Definition 3** Call a set of sections as a *domain*. Suppose we have two domains written by

$$D=\{d_i\} \ (i=1,\ldots,k) \text{ and } D'=\{d_j'\} \ (j=1,\ldots,l).$$

For D and D', OR, AND and COR operations are defined by:

$$D \text{ OR } D' = \{d_i \text{ OR } d_j' \mid \text{for all } i,j\}$$

$$D \text{ AND } D' = \{d_i \text{ AND } d_j' \mid \text{for all } i,j\}$$

$$D \text{ COR } D' = \{d_i \text{ COR } d_j' \mid \text{for all } i,j\}$$

**Example 1** For D={[2,3],[7,8,9,10]} and D'={[6,7,8],[12,13]}, we have:

D OR D' = {[2,3],[6,7,8,9,10],[12,13]}

D AND D' = {[7,8]}

D COR D' = {[6,7,8,9,10]}

**Theorem 1** For two domains D $\subset$ U and D' $\subset$ U, we have

D AND D' $\subset$ D COR D' $\subset$ D OR D' $\subset$ U.

*Proof* Trivial.

**Definition 4** For a domain D = $\{d_1, d_2, \ldots, d_k\}$, an extention operator ext(p)

is also defined by:

$$*D = ext(p)D = \{*d_1 \text{ OR } *d_2 \text{ OR } ..... \text{ OR } *d_k\} \quad (3)$$

Where the extended domain *D is a set of sections unified by OR operations for tuples of non-disjoint extended sections.

**Definition 5** Consider a text T and a short string K, termed *pattern*, such that #K<#T. If there exists a section d such that K=s[d], K is a *partial string* of T. We name the section d as a *place* of K. A procedure to detect the section d is called *pattern matching*.

**Definition 6** Consider a place d of pattern K in a text T (#T=n). Let *d be an extended section of d. Then, a string s[*d] is named as a *local context* of K. Here, *d is called a *field* of K.

**Definition 7** A set of all fields of pattern K in a text T is named a *contextual domain* of K. A set of all local contexts of K, corresponding to the fields, is named a *grobal context* of K. Suppose there exist k fields for a given pattern K. The contextual domain *D and grobal context C can symbolically be written by:

$$*D = \{*d_1, *d_2, ...., *d_k\}$$

$$C = \{s[*d_1], s[*d_2], ...., s[*d_k]\}$$

**4. Query operations**

Our purpose is to find optimal operations to access quickly the heart part in which the desired information is contained. Here, we discuss how to get to such heart parts using the operations previously introduced. In other words, this is nothing but how to carry out query operations on a given text.

Consider again the first of the three questions presented in the second chapter: Let T be a text of the bulky world history. We now have a hypothesis that the desired information is likely to be found in the common neighbourhood of given keywords; namely, what Caesar did in 60 BC may be described around sections belonging both to the contextual domain of pattern "Caesar" and to that of "60 BC". From this hypothesis, we can have the following procedure:

***Procedure 1***

I By pattern matching, detect all places of two patterns "Caesar" and "60 BC". Let $D_1$ and $D_2$ be the domains which contains all places of "Caesar" and "60 BC", respectively.

II Using the extension operator ext(p), create the two contextual domains $*D_1$ and $*D_2$ from $D_1$ and $D_2$, respectively.

III By AND or COR operation, a correlating part of $*D_1$ and $*D_2$ can be detected. Symbolically, we have

$$*D_1 \text{ COR } *D_2.$$

IV The desired information is expected to be found by reading carefully the strings on $*D_C$.

For $*D_A=*D_1$ AND $D_2$ and $*D_C$ appeared in III, obviously, we obtain $*D_A \subset *D_C$ from Theorem 1. Then, all strings on $*D_A$ are included by those on $*D_C$. This implies that when we finished to read all strings on $*D_C$ in advance, we can no longer obtain any more information from the strings on $*D_A$. In other words, strings on $*D_C$ are more redundant than $*D_A$, but more tolerant to loss of the information to be retrieved. Here, note that we can define a new domain $**D_A$ such that $*D_C \subset **D_A$ using the extention operator: By a proper q, we have

$$**D_A = ext(q)*D_A. \quad (4)$$

From this, we can also have a tolerant domain even if we did the AND operation. Anyway, it should be left to practical studies which operation, either AND or COR, is to be employed.

Procedure 1 provides a very optimistic strategy to retrieve what Caesar did in 60 BC. Consider the contextual domain $*D_1$ of "Caesar" appeared in II. In an actual text, "Caesar" is likely to be often substituted by the pronoun

"He" or "he". Sometimes, "Emperor" may be used instead of "Caesar". Then, it should be regarded as very optimistic to aim at building a retrieving strategy based on such $*D_1$ as defined simply by "Caesar". Obviously, we need to introduce more tolerant definitions introduced in our experiment.

**5. Experiment**

We already discussed query operations based on the hypothesis concerning to the contextual domains. Here, to testify to the hypothesis, query operations are experimentally performed on a very small-sized English text; *The Gospel According to Mark* of the New Testament. The reason why we have employed this text is that its content is so well-known to many people that our experimental results will properly be understood. Let this text be T (By our measurement, #T=74507).

Our experiment has been carried out to retrieve the exact strings to answer the following three queries:

Q1 "What did Jesus say on the Mount of Olives?"
Q2 "What happened between King Herod and the daughter of Herodia?"
Q3 "What did Jesus say at the Last Supper?"

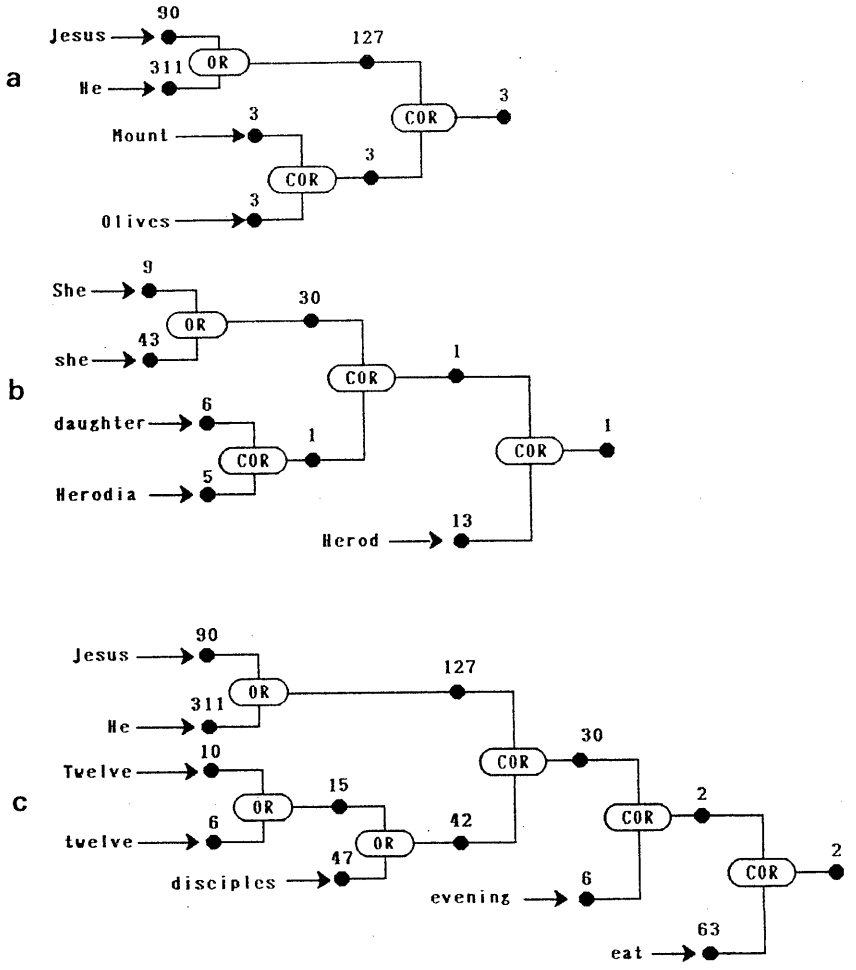


Figure 1. Retrieving trees for $Q_1$, $Q_2$ and $Q_3$.

Many people, especially Christians, probably have sufficient knowledge to answer these queries. Here, we consider how to build proper retrieving schemes using query oparations to obtain exact strings to answer Q1, Q2 and Q3. For experimental query operations, we have employed a constant extention operator set at p=100; i.e. ext(100). Our results and discussions are presented as follows:

**Q1**: This query contains the three important keywords, namely, "Jesus", "Mount" and "Olives". Among them, "Jesus" should carefully be handled: "Jesus" is so frequently replaced with "He" or "Him" that a preliminary operation is required to equalize them. In our experiment, two contextual domains of "Jesus" and "He" have been unified by OR operation.

Figure 1a shows our experimental scheme for the information retrieval. Call this type of figure as a *retrieving tree*. An arrow from each keyword to a black dot means the process of pattern matching, extention operation and defining the contextual domain. Black dots denote domains. A number written near a dot shows the number of detected sections included by the domain. Two domains are unified by an operation, encircled by a line, producing a new domain indicated in the right-hand side. All domains have eventually been unified into the rightmost *objective domain*, i.e. the root of the tree. If an operation produced the null domain, our retrieving tree would not been realized: In this case, either no objective domain exists in T, or our retrieving scheme is wrong or incomplete. In the latter case, sometimes, p is set at a too small value for the extention operator ext(p).

In Figure 1a, COR operation for two domains of "Mount" and "Olives" can be omitted when "Mount of Olives" is selected as a keyword. Thus, for a given query, the retrieving tree is not uniquely determined: Optimization of the tree will be required in practical situations.

Figure 1a illustrates that our objective domain contains three sections. The corresponding three strings are listed in Figure 2. On the top of each string, readers will find a numerical annotation such as 61.04~62.52: This means that the left and right ends of the section are 61.04% and 62.52% postions of #T, respectively.

Our result is satisfactory to Q1. Even if a string provides only a part of the words Jesus said, we can obtain the rest by applying the extention

----- 1 個目の影響領域要素です 61.04 ～ 62.52 -----

to be served but to serve and to give His life a ransom for many." 46 They came to Jericho and, as He was l
eaving Jericho with His disciples and a great throng, Bartimaeus, son of Timaeus, a blind begger, was sittin
g by the roadside. 47 Hearing that it was Jesus of Nazareth, he began to cry out, "Jesus, son of David, take
pity on me!" 48 Many ordered him to keep still; but he shouted the louder, "Son of David, take pity on me!"
49 Jesus stopped and said, "Call him." So they called to the blind man, "Have courage! Get up, He is callin
g you." 50 Throwing off his coat and springing to his feet he went to Jesus. 51 In response, Jesus said to h
im, "What do you want Me to do for you?" The blind man replied, "Rabboni, let me see again." 52 Jesus said t
o him, "Go! your faith has restored you." And instantly he recovered his sight and followed Him on the way.
CHAPTER 11 1 When they came near Jerusalem, at Bethphage and Bethany by the Mount of Olives, He dispatched t
wo of His disciples, 2 to whom He said, "Go to the village opposite you and, as soon as you enter it, you w
ill find a tethered colt w

----- 2 個目の影響領域要素です 74.18 ～ 75.22 -----

d from their surplus but she out of her poverty gave all she had-her whole living." CHAPTER 13 1 As He was l
eaving the temple, one of His disciples said to Him, "Teacher, see what wonderful stones and buildings these
are." 2 Jesus replied to him, "You see these great buildings? Not a stone shall be left on another, that sh
all not be torn down." 3 As He sat on the Mount of Olives opposite the temple, Peter, James, John and Andrew
asked Him privately, 4 "Tell us when this is to happen and what is the sign when all these things are to b
e accomplished!" 5 So Jesus began to tell them: "Look out that no one deceives you; 6 for many will come in
My name saying, 'I am He,' and will mislead many. 7 But when you hear about wars and rumours of wars, be not
alarmed; for it

----- 3 個目の影響領域要素です 83.56 ～ 83.87 -----

hen I shall drink it new in the kingdom of God." 26 With the singing of a hymn they went out to the Mount of
Olives. 27 And Jesus said to them, "You will all turn away from Me, for it is written, 'I shall strike the
shepherd and

Figure 2. Strings retrieved by the tree a in Figure 1.

operator to the section. This type of operation has also been employed in the case of Q2 and will be discussed in the following.

**Q2**: Important keywords characterizing this query include "Herod", "Herodia" and "daughter". Note that "daughter" is likely to be substituted by "She" or "she". Then, we first build a sub-tree including an OR and two CORs to detect the domains concerning to the daughter of Herodia. As seen in Figure 1b, only one section has been detected by this sub-tree. Eventually, we have obtained the objective domain which contains a single section: Figure 3a shows the string. Although the string is exact, it is too short to answer Q2. Then, by applying the extention operator ext(500) to the section, we have obtained a much longer string shown in Figure 3b. The latter string obviously provides sufficient information for Q2.

a

----- 1 個目の影響領域要素です 29.99 ～ 30.47 -----

on hearing him, was perplexed; yet he enjoyed listening to him. 21 An opportune time came when, on Herod's birthday, he gave a banquet to his nobles and commanders and prominent Galileans, 22 at which Herodias' daug hter came in and danced. She pleased Herod and his guests. So the king said to the girl, "Ask whatever you w ant and I will give it to you." 23

----- 1 個目の影響領域要素です 29.31 ～ 31.14 -----

b

of Him, Herod asserted, "John, whom I beheaded, has risen from the dead." 17 For Herod himself had sent to a rrest John and had confined him in prison, because of Herodias, his brother Philip's wife; for he had marrie d her. 18 For John had told Herod, "You have no right to have your brother's wife." 19 So Herodias held a gr udge against him and wanted to execute him but was unable to do so; 20 for Herod stood in awe of John becaus e he knew that he was an upright and holy man. He protected him and, on hearing him, was perplexed; yet he e njoyed listening to him. 21 An opportune time came when, on Herod's birthday, he gave a banquet to his noble s and commanders and prominent Galileans, 22 at which Herodias' daughter came in and danced. She pleased Her od and his guests. So the king said to the girl, "Ask whatever you want and I will give it to you." 23 Then he swore to her, "Whatever you ask me, I will give it to you up to half my kingdom." 24 She went out and ask ed her mother, "What shall I request?" "The head of John the Baptist," she replied. 25 She entered the hall and at once hastened to the king and made the request, "I want you to give me this moment on a platter the h ead of John the Baptist." 26 Although the King was extremely sorry, yet for the sake of his oaths and his gu ests he did not want to refuse her. 27 And at once the king di

Figure 3. (a)A string retrieved by the tree **b** in Figure 1. (b) The extended string by ext(500).

----- 1 個目の影響領域要素です 19.45 ～ 20.62 -----

ch large branches that the birds of the air can nest under its shelter." 33 With many such parables He told them the word insofar as they could grasp it; 34 He spoke in parables only and explained everything to His d isciples by themselves. 35 At evening that same day, He said to them, "Let us cross to the other side." 36 S o, leaving the crowd, they took Him along in the boat just as He was, and other boats accompanied Him. 37 A heavy squall of wind came up and the waves dashed into the boat so that the boat was filling, 38 while He wa s in the stern asleep on a pillow. They awoke Him and said to Him, "Teacher, do You not care that we are sin king?" 39 He rose up, rebuked the wind and said to the sea, "Silence! Be still!" Then the wind fell and ther e was great calm. 40 He said to them, "Why are you so afraid? Have you still no faith?" 41 They were terribl y frightened an

----- 2 個目の影響領域要素です 81.89 ～ 82.87 -----

wherever he enters, say to the proprietor, 'The Teacher says, "Where is My guest room where I am to eat the Passover with My disciples?"' 15 He will show you a large upper room-furnishings and everything ready-there prepare for us." 16 His disciples went out, came to the city and found it as He had told them. They prepared the Passover, 17 and as evening fell He arrived with the Twelve. 18 As they were sitting and eating, Jesus said, "I tell you with certainty that one of you who is eating with Me, shall betray Me." 19 They began to b e greatly disturbed, and they said to Him, one after another, "It is not I, is it?" 20 He answered them, "It is one of the Twelve, who is dipping with Me in the dish. 21 The Son of Man is g

Figure 4. Strings retrieved by the tree **c** in Figure 1.

**Q3**: This query is fairly complicated. No retrieving tree will be realized if it is unknown that "the Last Supper" is a picture painted by Leonardo da Vinci. Here, suppose the picture be known as depicting Jesus Christ and his

disciples at the supper table. Seven keywords have been selected in our experiment; namely, "Jesus", "He", "Twelve", "twelve", "disciples", "evening" and "eat". Figure 1c presents a retrieving tree composed by these keywords, producing the objective domain of two sections. The corresponding strings are shown in Figure 4. In this case, the first string is wrong, but the second is exact. Our retrieving tree looks to act properly, at least, in the sense that it brings no loss of information.

## 6. Conclusion

In this paper, emphasis has been placed both on introduction of the AND, OR and COR operations and on experimental works to examine their capabilities for information retrieval from a text. Our experiment shows that the proposed operations act very well in the information retrieval. In our experimental works, AND operation has not been employed at all, because AND operation associating with an extention operator can approximately be substituted by COR operation as discussed by (4). For each query to the Mark, a faily complicated retrieving tree has been constructed, which is needed for theoretical corroboration. In the practical situation, such a boring work should be left out: The user interface of a practical system will involve automatic translation of a given query into an optimal retrieving tree. For building the system, we will face technical problems including fast pattern matching. However, the recent innovation of computer technology suggests that we are to be very optimistic about such future problems.

*Aknowledgement:* The author thanks Mr Y. Kido and Mr T. Teranishi who helped the experimental works.

## References

[1]Boyer, R.S. and Moore, J.S., A Fast Searching Algorithm, *Comm.* ACM, Vol.20, No.11, pp 762-772, 1977.
[2]Aho, A.V. and Corasik, M.J., Efficient String Matching: An Aid to Bibliographic Search, *Comm.* ACM, Vol.18, No.6, pp 333-340, 1975.
[3]Uratani, N., FAST: A Fast Algorithm for Matching Multiple Patterns, Trans. IPSJ, Vol.30, No.9, pp 1119-1125, 1989.
[4]Luhn, H.P., Keyword-in-Context Index for Technical Literature (KWIC Index), American Documentation, Vol.11, pp 288-295, 1960.
[5]Smith, J.B., Weiss, S.F. and Ferguson, G.J., MICROARRAS: An Advanced Full-Text Retrieval and Analysis System, ACM SIGIR'87, pp 187-195, 1987.
[6]Gauch, S. and Smith, J.B., An Expert System for Searching in Full-Text, Inf. Process. Manage., Vol.25, No.3, pp 253-263, 1989.
[7]Raymond, D.R. and Fawcett, H.J., Playing Detective with Full Text Searching Software, ACM SIGDOC Asterisk, Vol.14, No.4, pp 157-166, 1990.
[8]Grefenstette, G., Use of Syntactic Context to Produce Term Association Lists for Text Retrieval, ACM SIGIR'92, pp 89-97, 1992.