Approach towards realizing resource mining and secured information transfer

Prasun Chakrabarti , Partha Sarathi Goswami
Dr. B.C.Roy Engineering College , Durgapur-713206,West Bengal, India

Summary
The paper deals with the concept of data mining whereby the data resources can be fetched and accessed accordingly with reduced time complexity. The retrieval techniques are pointed out based on the ideas of binary search tree, Gantt chart, text summarization. Summarization is a hard problem of Natural Language Processing because, to do it properly, one has to really understand the point of a text. This requires semantic analysis, discourse processing, and inferential interpretation (grouping of the content using world knowledge). The last step, especially, is complex, because systems without a great deal of world knowledge simply cannot do it. Therefore, attempts so far of performing true abstraction--creating abstracts as summaries--have not been very successful. Fortunately, however, an approximation called extraction is more feasible today. To create an extract, a system need simply to identify the most important/topical/central topic(s) of the text, and return them to the reader. Although the summary is not necessarily coherent, the reader can form an opinion of the content of the original. Most automated summarization systems today produce extracts only. Lastly, extraction of resources can be efficiently done using statistical approaches. Another aspect of the paper is concept of shared key in case of multiparty communication. Information security plays a pivotal role. Various proposed techniques have been cited for key evolution in multi-party domain and the areas concerned are based on interlock protocol , SKEY and SKID.

Key words : *data mining , time complexity, binary search tree , Gantt chart, text summarization ,interlock protocol, SKEY,SKID*

1. Introduction
Accessing information that is resources from heterogeneous data should be done in an optimum way. The search tree can be applied for effective search. The average waiting time for successful transaction of data can easily be analyzed with the help of Gantt chart whereby we denote search transaction for an user as a process. Sometimes in case of web mining of resources, the context of text summarization is done where the search is based on some selected portion of text. Herein lies the importance of text summarization which is based on centroid-based algorithm. Another way of retrieval is based on statistical approached where prediction of data is the main factor and also time management should be in an optimum fashion. Information security plays a pivotal role in case of data transfer. Variability of key can be applied instead of fixed key concept and the data transfer can be done based on public key cryptosystems.

2. Mining of resources
A search can be formed based on the initial search term and its gradual sub term while the process of matching. Thereby the level is increased, in initial search term is the root and the final term fully matching with the context of the users' desire is a leaf node.

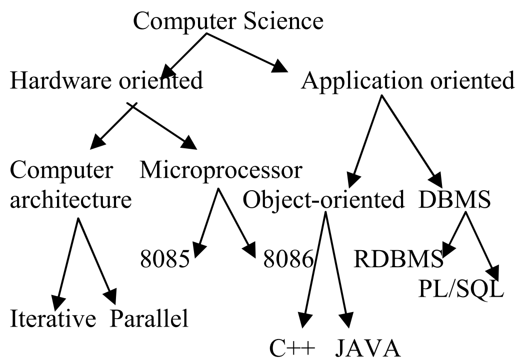

Fig1: Binary search tree

In the above figure, Computer Science is the root that is initial search term. If a user wants to access resources PL/SQL, then the database hierarchy, PL/SQL is the node in level 3 (initial level is 0) and it is a leaf node. For future purpose if the database administrator saves the model in a database and identify each search term as a binary code, then by giving the code number he can analyze the position of data in the model and acknowledge quickly as per users' request. The concept of coding is as follows:

Value = 0 if the search term is a left child of parent node
= 1 otherwise

N

Manuscript received July 5, 2008
Manuscript revised July 20, 2008

2.1. Theorem: In the process of coding, $\sum_{i=1} 1/2^{L_i} = 1$, where L_i is the length of code of ith leaf node in the tree, N is total number of leaf nodes and 1<i<N.

2.1.1. Proof:

In Fig. 1 codes of leaf nodes are as follows:
Hieratic Architecture :000
Parallel Architecture : 001
8085 model : 010
8086 model : 011
C++ : 100
JAVA : 101
RDBMS: 110
PL/SQL " 111
So, N=8. Each leaf node has identical code length i.e. 3.
Therefore, $1/2^{L_i}$ =1/2=1/8, 1/2= 1/8, …1/2=1/8

3. Mining of the resources based on Gantt chart

Let R be a resource and the users are A, B, and C. Now, for transaction of R , each of A, B and C send request to the database administrator. According based on the priority, the search schedule is performed.
Let, P1 = Process of search by user A
P2 = Process of search by user B
P3 = Process of search by user C

Let, tP1 = time for A to search successfully = 4 seconds
tP2 = time for B to search successfully = 6 seconds
tP3 = time for C to search successfully = 3 seconds
If priority of P1>P3>P2, then Gantt chart is s follows:

P1	P3	P2

0 4 7 13

Waiting time of P1 = 0, waiting time of P2 = 7 seconds and waiting time of P3 = 4 seconds.
Sometimes the concept of Round Robin Scheduling is applied whereby a time slice is given and after that the process is switched to another user irrespective of completion time of search. Let, time slice = 2 seconds, then the Gantt chart is as follows:

P1	P3	P2	P1	P3	P2	P2

0 2 4 6 8 9 11 13

Hence, after 9th seconds , two successive search engines are performed by user B as the other users A and C have already fetched their information successfully.

4. Mining of the resources based on centroid based text summarization

The mining technique is based on Centroid-based algorithm which is as follows :
Input**:** A collection of related documents.
Output**:** A summary.

Steps to summarize :

4.1. Finding Cluster Centroid
A cluster consisting of total number of sentences from all input documents is formed. The 'count' value for each word indicating the average number of occurrences of a word across the entire cluster is found out. Then the centroid value for each term is calculated as:

Count * idf(w)=count(w) * (log(DN / df(w)))

where df(w)=document frequency for each word.
DN=number of documents in the corpus.

4.2. Finding Sentence Position Score
The score of ith sentence **(Si)** is computed
Pscore*(Si)*= max(1 / i , 1 / (n-i-1))
where i=sentence number
n=number of sentences

4.3. Finding Sentence Length Score
The length here means the number of characters in the sentence. A sentence shorter than a certain length gets penalty. The length score of a sentence can be calculated as
Lscore*(Si)*= 0 if Li≤ Lmin
= (Li-Lmin) / Li otherwise
where Li=length of each sentence
Lmin=20 , i.e. sentence with 20 or fewer characters receives penalty.

4.4. Finding Headline Score
The idea is that greater the number of words in a sentence that match those in the headline , the more important the sentence is likely to be. The headline score can be calculated as

Hscore*(Si)*= t / N

where t=number of words in the sentence that match with the words in the headline
N= number of words in the sentence

4.5. Compute Sentence Score

SCORE*(S)*=∑ (wc.Ci + wp.Pi + wf.Fi + wl.Li)

where i ranges from 1 to n as (1≤i≤n)

Also, Ci=Centroid value of the sentence
Pi=sentence position score
Fi=headline score
Li=sentence length score

Wc= wI = wf = wl =1
n = number of sentences in the cluster

4.6. Extract Sentences

Sentences are sorted according to descending order. Select d out of n sentences as an intermediate summary of the input documents. The sentences are extracted in an order.

$$d = r * n$$

where r = Compression Rate
and n = total number of sentences taken from input documents.

5. Security implementation using public cryptosystems

5.1. Interlock protocol in the light of variable key :

1) Alice and Bob generate a session key for sharing. Let $K_{AB.}$
2) Alice encrypts its public key and sends $E_{KAB}(K_{APUBLIC})$ to Bob. Bob sends $E_{KAB}(K_{BPUBLIC})$ to Alice.
3) Alice decrypts and gets $K_{BPUBLIC}$. She then sends half of the message for Bob in encrypted form by $K_{BPUBLIC}$.
4) Similarly Bob does so.
5) Alice then computes $K_{APUBLIC}$' = Modification of $K_{APUBLIC}$, sends $E_{KAB}(K_{APUBLIC'})$ to Bob.
6) Similarly Bob does so.
7) Alice then sends $E_{KBPUBLIC'}$(other half of message) to Bob.
8) Similarly Bob performs.
9) Each receiver then decrypts the message in parts by respective keys and retrieve the message sent to him/her.

5.1.1. Mathematical Analysis

5.1.1.1. Encryption

Let the session key for sharing be the binary form of date. Let the date be 10.06.08 then: Binary form of 10 is 1001
Binary form of 6 is 0110
Binary form of 8 is 1000.
Perform XOR on 1001 and 0110 it gives 1111. The perform the next XOR 1111 with 1000 it gives 0111. The decimal form of 0111 is 7. This is the session key.

Again Let us take a super increasing knapsack sequence, for example {2, 3, 6, 13, 27, 52}, and multiply all of the values by a number *n,* mod *m*. The modulus should be a number greater than the sum of all the numbers in the sequence: for example, 105. The multiplier should have no factors in common with the modulus: for example, 31. The normal knapsack sequence would then be

2 * 31 mod 105 = 62
3 * 31 mod 105 = 93
6 * 31 mod 105 = 81
13 * 31 mod 105 = 88
27 * 31 mod 105 = 102
52 * 31 mod 105 = 37

The knapsack would then be {62, 93, 81, 88, 102, 37}. The super increasing knapsack sequence is the private key. The normal knapsack sequence is the public key. Let the first half of the message be 110011 in binary form encryption using the previous knapsack would proceed like this:

Message = 110011 corresponds to 62+93+102+37=294.. The value of Session Key=7.New Message= Old message – 7=294 – 7=287.The cipher text would be 287,7.

5.1.1.2. Decryption

The super increasing knapsack is {2, 3, 6, 13, 27, 52}, *m* is equal to 105, and *n* is equal to 31. The cipher text message is 287,7. In this case n^{-1} is equal to 61, so the cipher text values must be multiplied by 61 mod 105.Original Cipher text=287 + 7=294. Now 294 * 61 mod 105 = 14 =1+2+5+6, which corresponds to 110011.The recovered first half of plaintext is 110011.

5.2. Analyzing of SKEY in the light of variable key :

SKEY is mainly a program for authentication and it is based on a one-way function.
The steps are as follows:

1) Host computes a Bernoulli trial with biased coin for which p= probability of coming 1, q=(1-p)=probability of coming 0.Let number of trials be n. Assume n=6, and string=110011.
2) Host sends the string to Alice.
3) Alice modifies its own public key based on that the new public key = previous key + (binary equivalent of the number of 1's present in the string).
4) Alice creates a Shared Key.
5) Alice modifies the public key along with modification scheme with shared key.

6) Alice then encrypts the string with her private key and sends back to the host along with her name.
7) Host first decrypts public key and accordingly fetches it from database of Alice and computes the result.
8) If match is found, then it performs another level of verification by decrypting the string with new value of Alice's public key.
9) If that also matches, then authentication of Alice is certified.

5.2.1. Mathematical Analysis

5.2.1.1.Encryption

Let us take a super increasing knapsack sequence, for example {2, 3, 6, 13, 27, 52}, and multiply all of the values by a number *n,* mod *m*. The modulus should be a number greater than the sum of all the numbers in the sequence: for example, 105. The multiplier should have no factors in common with the modulus: for example, 31. The normal knapsack sequence would then be

2 * 31 mod 105 = 62
3 * 31 mod 105 = 93
6 * 31 mod 105 = 81
13 * 31 mod 105 = 88
27 * 31 mod 105 = 102
52 * 31 mod 105 = 37

The knapsack would then be {62, 93, 81, 88, 102, 37}. The super increasing knapsack sequence is the private key. The normal knapsack sequence is the public key. If the message is 110011 in binary, encryption using the previous knapsack would proceed like this:

Message = 110011 corresponds to 62+93+102+37=294
No. of 1's=4, Binary form of 4=100
New Message= Old message + (no. of 1's in binary form)=294 + 100=394

The cipher text would be 394,4.

5.2.1.2. Decryption

The super increasing knapsack is {2, 3, 6, 13, 27, 52}, *m* is equal to 105, and *n* is equal to 31. The cipher text message is 394,4. In this case n^{-1} is equal to 61, so the cipher text values must be multiplied by 61 mod 105.

Original Cipher text=394 – (Binary form of 4)=394 – 100= 294. Now 294 * 61 mod 105 = 14 =1+2+5+6, which corresponds to 110011.The recovered plaintext is 110011

5.3. Analysis of SKID in the light of variable key:

The steps are as follows:

1) Alice chooses a random number R_A and sends it to Bob.
2) Bob chooses a random number R_B and sends it to Alice.
3) Alice and Bob make a secret shared key K.
4) Bob generates $R_{A'}$, $R_{B'}$, K' and sends $E_K(R_{A'}, R_{B'}, K')$ and $H_K(R_{A'}, R_{B'}, B)$ to Alice, H_K being for the MAC.
5) Alice extracts R_A, R_B, K and then computes $H_K(R_{A'}, R_{B'}, B)$ to find B. Then she matches that with what was sent to her by Bob.
6) If match= true, Alice knows she is communicating with Bob.

5.3.1. Mathematical Analysis

Let R_A = some prime no.= p = 3 (say)
R_B = some other prime no.= q = 5 (say)
K= p * q = 3 * 5 = 15
Let $R_{A'}$ = ln(p)= ln(3)
$R_{B'}$ = ln(q)= ln(5)
K' = ln(K)= ln(15)
$E_K(R_{A'}, R_{B'}, K')$ = ln(3) * ln(5) * ln(15) = ln(225) and
$H_K(R_{A'}, R_{B'}, B)$ = ln(3) * ln(5) * ln(25) = ln(375),
Let B=MAC = 25.
This is $E_K(R_{A'}, R_{B'}, K')$ and $H_K(R_{A'}, R_{B'}, B)$ is being send to Alice.

Alice extracts by $DE_K(R_{A'}, R_{B'}, K') = e^{EK} = e^{\ln(225)} = 225 = 3 * 5 * 15 = R_A * R_B * K$
$DH_K(R_{A'}, R_{B'}, B) = e^{HK} = e^{\ln(375)} = 375 = 3 * 5 * 25 = R_A * R_B * B$
Since match= true, Alice knows she is communicating with Bob.

6. Statistical approaches of resource mining

6.1. Based on prediction of most frequent word:

The most frequent word can be obtained based on $Max(f_1,f_2,\ldots,f_n)$ where f_1, f_2, ….,f_n are relative frequencies and n is total no. of words.

6.2. Based on prediction of variable within interval:

We can predict the value of a variable if we can measure interval properly. We can apply this scheme in hacking.

6.2.1. Theorem:

If a variable changes (V) over time (t) in an exponential manner, in that case the value of the variable at the centre point an interval (a_1, a_2) is a geometric mean of its value at a_1 and a_2.

Proof: Let $V_a = mn^a$

Then $V_{a1} = mn^{a1}$ and $V_{a2} = mn^{a2}$

Now, value of V at $(a_1 + a_2)/2$

$$= mn^{(a1+a2)/2}$$
$$= [m^2 n^{(a1+a2)}]^{1/2}$$
$$= [(mn^{a1})(mn^{a2})]^{1/2}$$
$$= (V_{a1}V_{a2})^{1/2}$$

6.3. Based on prediction of interrelated variables:

In a message there may be a variable which is dependent on any other based on any equation in that case extraction can be made.

6.3.1. Theorem:

If a variable m related to another variable n in the form m = an, where a is a constant, then harmonic mean of n is related to that of n based on the same equation.

Proof: Let x is no. of given values.

If $m_{HM} = x / (\sum 1/m_i)$ for i = 1 to x

$= x / (\sum 1/an_i)$ [Since $m_i = an_i$]

$= x / (1/a \sum 1/n_i)$ for i = 1 to x

$= a(x / (\sum 1/n_i)$ for i= 1 to x

$= an_{HM}$

7. Conclusion

In the paper we have observed that efficient ways of optimum data mining reduces time complexity. In case of text summarization. Researches are going on this topic. There are many other techniques related to text summarization based on position of sentences or length of sentences of the documents. It will be more reliable if the sentences are parsed in phrase level using Link Grammar parser. For each sentence with the content of the sentence there should should be associated the information of the words of the sentence. The information of the word means ‘subject’, ‘time’, ’space/ location’, ‘action i.e. verb’ etc. Using these information the sentences are clustered on the basis of same ‘subject’ or ‘action’ etc. These clusters are ranked on the basis of size. The clusters are extracted from top order until required summary length is achieved. The estimated approaches of resource mining has been pointed out based on prediction in the light of statistical approach. It has also been shown how variable key can be applied efficiently in case of interlock protocol , SKEY and SKID in case of secured data transfer in multi-communication domain.

References

[1] Daniel N ; Radev D, and Allison T (2003) - Sub-event based multidocument summarization. In HLT-NAACL Workshop on Text Summarization, Edmonton, AB, Canada

[2] Grewal A ; Allison T ; Dimitrov S and Radev D (2003) - Multi-document summarization using o_ the shelf compression software. In HLT-NAACL Workshop on Text Summarization, Edmonton, AB, Canada

[3] Kareem O and Radev D (2004) - Hierarchical text summarization for WAP-enabled mobile devices. Submitted to SIGIR 2004 Demo Session

[4] Otterbacher J and Radev D (2004) - A resource for revision-based multi-document summarization and evaluation. In LREC, Lisbon, Portugal

[5] Radev D; Blair-Goldensohn S, and Zhang Z (2001) - Experiments in single and multi-document summarization using M EAD. In First Document Understanding Conference, New Orleans, LA

[6] Chakrabarti P. and et.al. (2006) , “Centroid Based Multidocument Summarization: efficient sentence extraction method” published in 26th Annual International conference of South Asian Language Analysis (SALA26), jointly hosted by Kannada University, Hampi and Central institute of Indian Languages, Mysore, 19-21Dec 2006

[7] Chakrabarti P. and et.al. (2007) , “Effective news corpus and centroid based extraction” , published in the International Conference on “Ethnographic discourse of the other in contemporary India” at University of Hyderabad , July 07

[8] Chakrabarti P. and et.al. (2007) , “Shared key evaluation in multiparty communication” published in International Conference on IT, Jabalpur, Dec07

[9] Chakrabarti P. and et.al. (2007) , “Sentence extraction scheme in Centroid based multi-document summarization and an approach of News Corpus development “published in State level National seminar at B.B.College, Asansol on 16-17th Mar07

[10] Schneir Bruce (1996),“Applied Cryptography”, John Willey and Sons Inc., New York, USA

[11] Giri P.K. and Banerjee J. (1999) ,”Introduction to Statistics”, Academic Publishers , India

ABOUT AUTHORS

Prasun Chakrabarti (09/03/1981) has received his Bachelor of Technology (B.Tech) degree in Computer Engineering from B.P. Poddar Institute of Management and Technology under Kalyani University, India in 2003 and Master of Engineering (M.E.) degree in Computer Science and Engineering from West Bengal University of Technology, India in 2005.He is now pursuing his Ph.D(Engg.) in the area of Information Security from the department of Computer Science and Engineering , Jadavpur University, India .His research areas include information security , data mining , fuzzy systems, etc. He has about 12 research papers in national/international journals and 30 research papers in national/international conferences in his credit. He is a professional member of IEEE , Indian Science Congress Association , Calcutta Mathematical Society , Calcutta Statistical Association , Indian Society for Technical Education ,Computer Society of India, VLSI Society of India, Institution of Electronics and Telecommunication Engineers , Operation Research Society of India , Cryptology Research Society of India , Society for Information Science , Forum of Scientists and Technologists. He has also several teaching and research honors. He is a reviewer of the International Journal of Information Processing and Management (IPM) , USA.

Partha Sarathi Goswami (27/04/1979) has received Bachelor of Science (B.Sc.) degree in Mathematics (Honors) from Moulana Azad College (Govt. of West Bengal) under University of Calcutta, India in 2000, his Master of Computer Applications (MCA) from RCC Institute of Information Technology under University of Kalyani, India in 2003 and Master of Philosophy (M.Phil) degree in Computer Science from Annamalai University, India in 2007. He is now in his final semester in Master of Technology (M.Tech.) in Computer Science and Engineering at UPTU, India. He has more than 5 years of teaching experiences in different Engineering and Technology colleges in India. He has guided many students of MCA, BCA, B.Tech. and MBA in their project work. His research areas include information security , data mining , fuzzy systems, computer graphics, etc. He has 2 research papers in international journals.