Doctoral Committee Meeting No. 2

Web Search Engine

Sanasam Ranbir Singh (CS05D018)

Under the Guidance of

Prof. Hema A. Murthy

Prof. Timothy A. Gonsalves

Contents

- **Web Search:**
  - **Not only Document-Content Search**
- **Literature Survey**
- **Progress**
- **Future Works**

Web Search:
Not only Document-Content Search

Query ?

Web Search:
Not only Document-Content Search

Query ?

Cancer Hospitals at Chennai ?

Web Search:
Not only Document-Content Search

Query ?

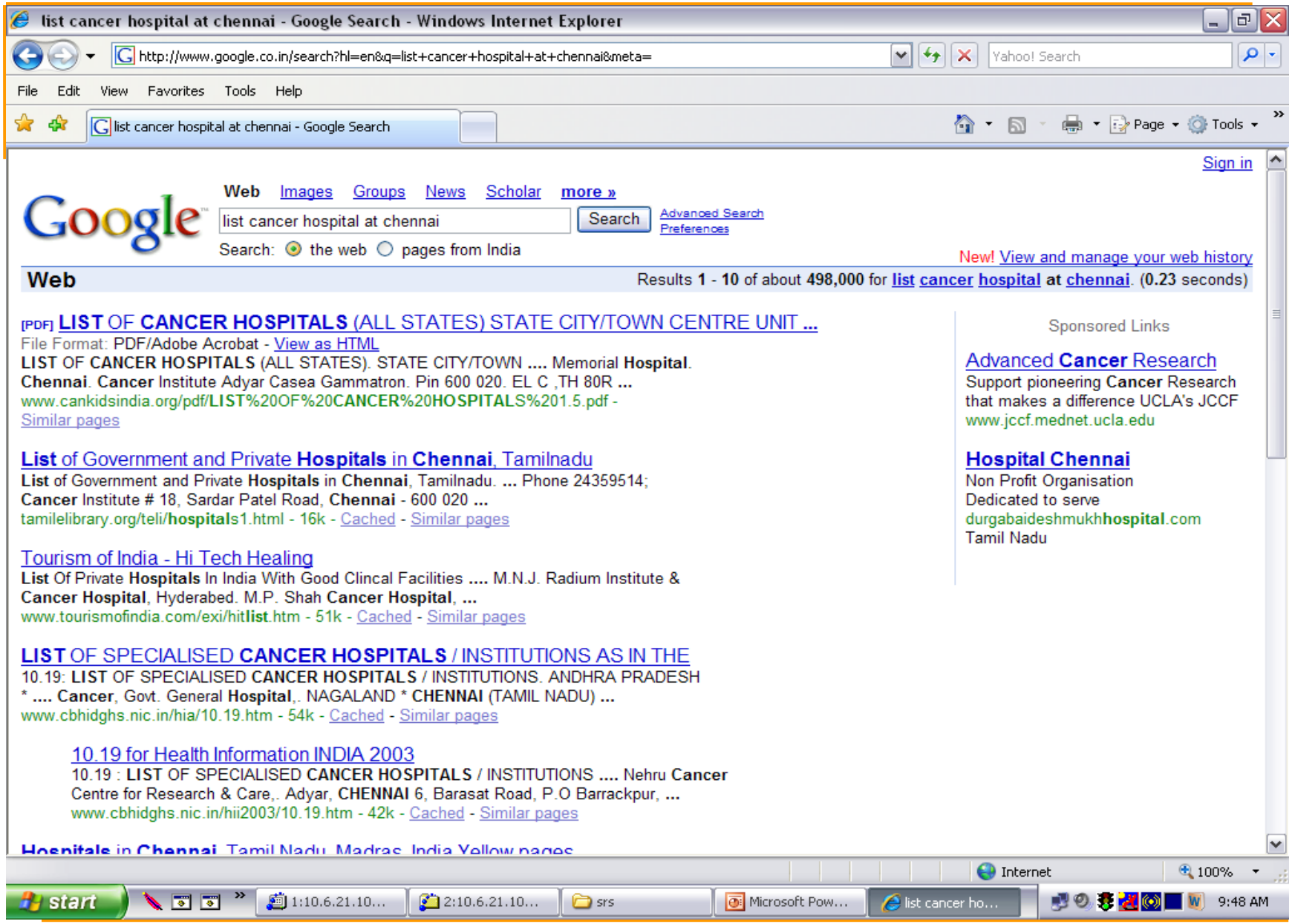
Cancer Hospitals at Chennai ?

OR

Best Cancer Hospitals at Chennai ?

Web Search: Not only Document-Content Search

- General search engines find documents containing the query terms



S. Ranbir Singh (2nd DC, 24th Sep.-2007)

Web Search:
Not only Document-Content Search

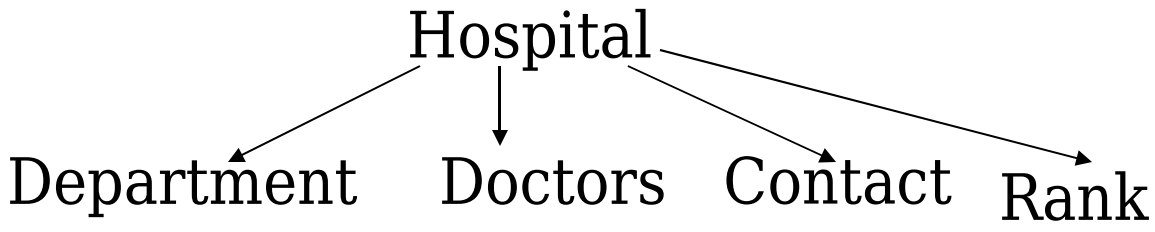
- General search engines find documents containing the query terms

- Instead, Users want the Search Engines to find the hospitals, their details and present the summary to the users

Associated Words and Its Applications

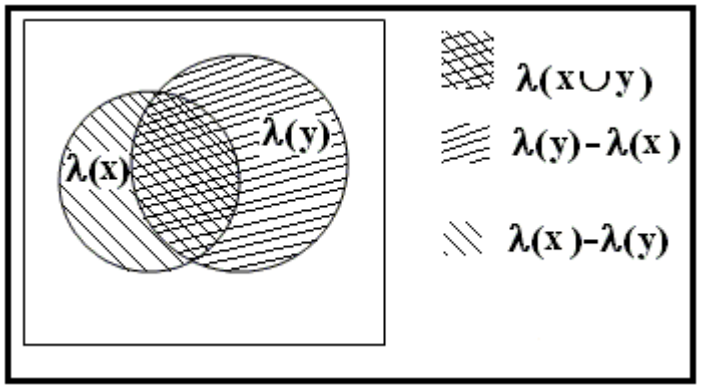
What are Associated words?

Cancer to prostate, oncology, tumor, …. etc.



Literature Survey

- Words Association Norms [1]
  - p(x,y)/p(x).p(y)
- Mutual Information [2]
  - Expectation of Words Association Norms
- Association Rules [3]
  - X describes Y (X →Y)



Progess

- Association rules and Joint Relative Coverage Estimators

$$Score(X,t) = c(t \rightarrow X)\left(\frac{\sigma(t \cup X, \mathcal{D})}{\sigma(t,\mathcal{D}) + \sigma(X,\mathcal{D}) - \sigma(t \cup X,\mathcal{D})}\right)c(X \rightarrow t)$$

$$= c(t \rightarrow X)\left(\frac{p(t,X)}{p(t) + p(X) - p(t,X)}\right)c(X \rightarrow t)$$

- Effect of Words density on Words Association
  - Consider the words density in account
- A greedy approach of Clustering Data Points
  - For every data point x, identify another data point y satisfying the property that wherever y goes, x follows.

Results

Dataset

Characteristics of the documents in the repository. We consider the words with length≥ 3 only. This repository is collected from Web during the month of October 2006 using WIRE web crawler.

Number of sites involved	20000
Number of documents downloaded	496114
Number of urls found in the documents	4288261
Number of unique words in the repository	1895184
Average number of unique words appeared in a document	230.171
Standard deviation of the # of unique words appear in a document	276.745
Minimum number of unique words appeared a document	5
Maximum number of unique words appeared in a document	8871
Average support of a word	49.748
Standard deviation of the support of a word	127.14
Minimum support of a word	1
Maximum support of a word	300,145

15 Terms

Words density not consider

Precision:

How well one describes the other?

Use

-Google, AltaVista, Askc

Comparison of the average precisions of results returned by different estimators. The θ^* in ARJP estimator means, it has considered $\Theta^{conf} = 90\%$

estimator	θ	AvgPre10	AvgPre20	AvgPre30	AvgPre40	AvgPre50
co-occurrence	2	0.136	0.122	0.1202	0.1159	0.1102
	3	0.1945	0.1622	0.1665	0.1664	0.1653
	4	0.1961	0.1663	0.1758	0.1722	0.1752
	5	0.4509	0.46	0.4526	0.4572	0.4534
Mutual Information	2	0.251	0.2106	0.1925	0.1848	0.1712
	3	0.3703	0.3522	0.3539	0.3492	0.3411
	4	0.3751	0.365	0.3662	0.3619	0.3649
	5	0.451	0.4646	0.467	0.4687	0.4673
Association Rule	2	0.3118	0.256	0.2215	0.2002	0.1832
	3	0.5968	0.5791	0.5647	0.5586	0.5524
	4	0.6554	0.6554	0.6454	0.6376	0.6331
	5	0.6925	0.7124	0.7107	0.7005	0.6987
Association Rules with Joint Probability	2	0.2517	0.2397	0.2139	0.1976	0.1705
	3	0.5492	0.5402	0.5273	0.5172	0.5108
	4*	0.6201	0.6144	0.6004	0.5863	0.5751
	4	0.5828	0.5937	0.6025	0.598	0.6064
	5	0.5027	0.5124	0.5041	0.4956	0.5004
Association Rule with Joint Relative Coverage	2	0.3191	0.2464	0.2297	0.1955	0.1875
	3	0.5912	0.5859	0.5596	0.5547	0.56
	4	0.6642	0.6747	0.6682	0.6684	0.6663
	5	0.6927	0.7025	0.7145	0.7083	0.7068

Effect of Words Density on Words Associations

Comparison of the average precisions of results returned by different estimators. The full documents are taken as the windows.

estimator	AvgPre10	AvgPre20	AvgPre30
Joint Probability	0.129	0.123	0.123
Association Rule	0.308	0.262	0.235
$R'(x,y)$	0.31	0.264	0.237
$R(x,y)$	0.496	0.487	0.482

R/(x,y): ARJRC

R(x,y): Incorporate Words Density

Window: sequence of words

Comparison of the average precisions of results returned by different estimators. The size of a window is the size of a sentence.

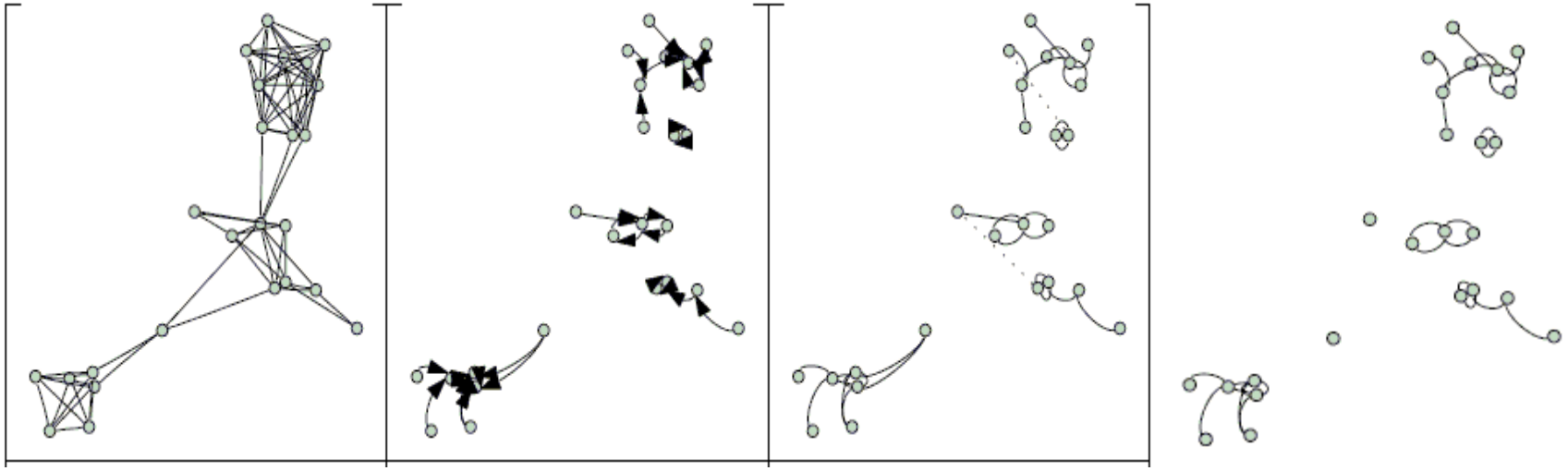
estimator	AvgPre10	AvgPre20	AvgPre30
Joint Probability	0.317	0.31	0.298
Association Rule	0.494	0.492	0.488
$R'(x,y)$	0.496	0.491	0.487
$R(x,y)$	0.514	0.51	0.506

Comparison of the average precisions of results returned by different estimators. A window is a collection of five to ten sentences.

estimator	AvgPre10	AvgPre20	AvgPre30
Joint Probability	0.427	0.511	0.513
Association Rule	0.562	0.561	0.556
$R'(x,y)$	0.554	0.551	0.549
$R(x,y)$	0.571	0.567	0.565

A greedy approach of Clustering Data Points

Wherever y goes, x follows.



Graph Split

Initial Cluster

Merge

Publications

Published

- Estimating the Rate of Web Page Updates, S. Ranbir Singh, 12th IJCAI, 2007

Communicated

- Mining Related Terms, S. Ranbir Singh, Hema A. Murthy, Tag A. Gonsalves, DKE, Elsevier.
- Effect of Words Density on measuring Words Association, S. Ranbir Singh, Hema A. Murthy, Tag A. Gonsalves, ACM Compute 2008

Future Works

1. Clustering algorithms for possible publication.
2. Exploring more on Associated Words
3. Integrating multi-lingual Issues on Information Retrieval

References

[1] K. W. Church, P. Hanks, Words association norms, mutual information and lexicography, Computer Linguist. 16(1), 1990 22-29.

[2] M. T. Tomohiko et la. A method of extracting related words using standard Mutual information. Discovery Science. 2843 (1), 2003 .

[3] R. Agrawal, R. Srikant, Fast algorithms for mining association rules. 20th Int. Conf. on Very large Data Bases, 1994.