Digital Library Curriculum Development

Module 7-d: Routing

1. **Module Name:** Routing

2. Scope

This module presents an overview of models and practices of Routing, leading to better search result quality in Digital Libraries.

3. Learning objectives

Students will be able to:

1. Explain the need for routing in digital libraries
2. Explain the basic concepts of routing
3. Identify the different types and methodologies of routing systems
4. Explain the challenges involved in routing techniques

4. 5S characteristics of the module

- **Stream:** Routing systems primarily deal with textual information. But they also may handle media types such as images, audio, and video.
- **Structure:** Routing requires particular structures for queries, profiles and forward knowledge systems.
- **Space**: In routing, vector space finds the similarity of query and document vectors.
- **Scenario**: The process of routing the query to the identified appropriate digital reference services.
- **Society**: End users (such as librarians, computer programmers, journalists) requiring information and systems (such as routers) that facilitate the retrieval.

5. Level of effort required

a. In class: 4 hours
b. Outside of class:
    i. 2 hours for readings
    ii. Approximately one hour for the homework assignment (refer to Exercises section below)

6. Relationship with other modules

Close connections with:

- 6-a: Info needs, Relevance - Module 6-a is a prerequisite to Module 7-d
- 7-b Reference Services - Module 7-b should be taught before module 7-d

7. Prerequisite knowledge required

- General understanding of library reference services
- A prior basic understand of internet infrastructure

8. Introductory remedial instruction

None

9. Body of knowledge

I. Need for Routing

With the rapidly growing, heterogeneous, distributed collection of data sources available in the Internet, locating and accessing information has been a significant problem. This is mainly applicable for large-scale digital libraries. This problem is made worse by the rapid growth in information in the web and by the increasing number of naive users who typically issue very short or broadly defined queries to search systems. Query routing techniques offer an important advantage of facilitating focused resource discovery. This enables the queries to focus on certain semantic characteristics of the resource and the discovery focuses on a specific category of resources.

II. Introduction to Routing

Query routing is used to redirect the queries from the user to the appropriate digital reference services. The main idea of query routing is to restrict the scope of the search through various methods, thus making it effective. An efficient query routing system minimizes the response time for fetching the results by reducing unnecessary communication overhead over the network to the individual digital reference services. So, the main goal of query routing is to route a user query to the most relevant information sources that can provide the best answer.

III. Types of Routing Systems:

A. Query Routing

i) Manual query routing services

General-purpose search engines (such as AltaVista, Lycos) are known for returning non-relevant search results for the user query. This led to the demand for topic specific search. Some search engines (such as VacationSpot.com, KidsHealth.org etc.) provide a categorized list of specialized search engines for particular topics. Conventional query routing services to specific search engines are done manually by the user. The users themselves are required to choose appropriate search engines from the list based on the information need.

ii) Automated query routing systems

There are certain routing systems which perform automated query routing. Some of these systems use content summaries for categorizing the databases based on the list of term-frequency pairs. With these summaries, the systems can identify if a particular database is relevant to the user query or not.

B. Document Routing

In document routing, the file or the document is routed from one user to another in a network. This type of routing is quite similar to query routing except that the content which gets transferred is a document in this case. The methods and algorithms used to route the content will mostly be the same in all the cases. This type of routing usually happens in a peer-to-peer file sharing network, where one peer tries to share a file to another node across the network.

C. Selective Dissemination of Information (SDI) Systems:

With the wealth of information available in the internet, the volume of information available ranging across all interests has also increased. As already explained in the need for query routing, it is difficult to query and filter formation pertaining to individual interests. The aim of an SDI system is to deliver new information arriving at an SDI-aware information provider to users who express their interests through user profiles.

There are various ways of implementing SDI systems - one such option being persistent query mechanisms, where the users create a query and add them to the system. As the queries remain resident in the system, successful matches with incoming documents to the information sources are identified and routed to the particular user.

Content Routed	Routing Type
user query	Query Routing
file / document	Document Routing
information of user's interest	Selective Dissemination of Information (SDI)

IV. Query routing architecture

A. Mediator based routers:

A query routing system could have any structure depending on network topology. One such system looks like a directed acyclic graph with router nodes present in the middle and the information providers are the leaf nodes. This hierarchy is shown in figure-1. The router nodes usually contain the content summary and query capability description of the information providers that are registered with it. In some methods, the role of query routers is done by the mediator, which contains forward knowledge about the digital reference services.

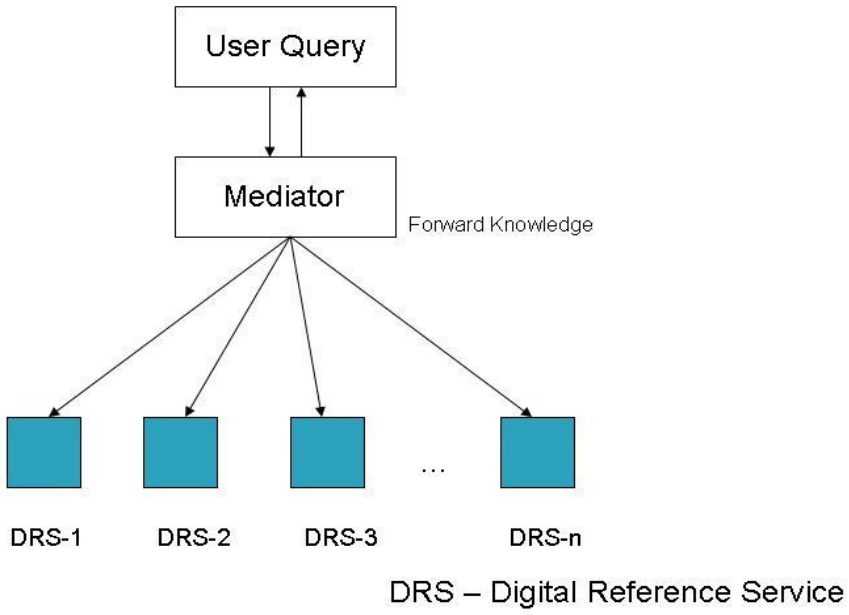


Figure-1 Illustration of a mediator based routing architecture

B. Peer to Peer based routers:

In the case of peer to peer methods, the routing is done by each peer in the network. One good example is the Gnutella network which is shown in figure-2.

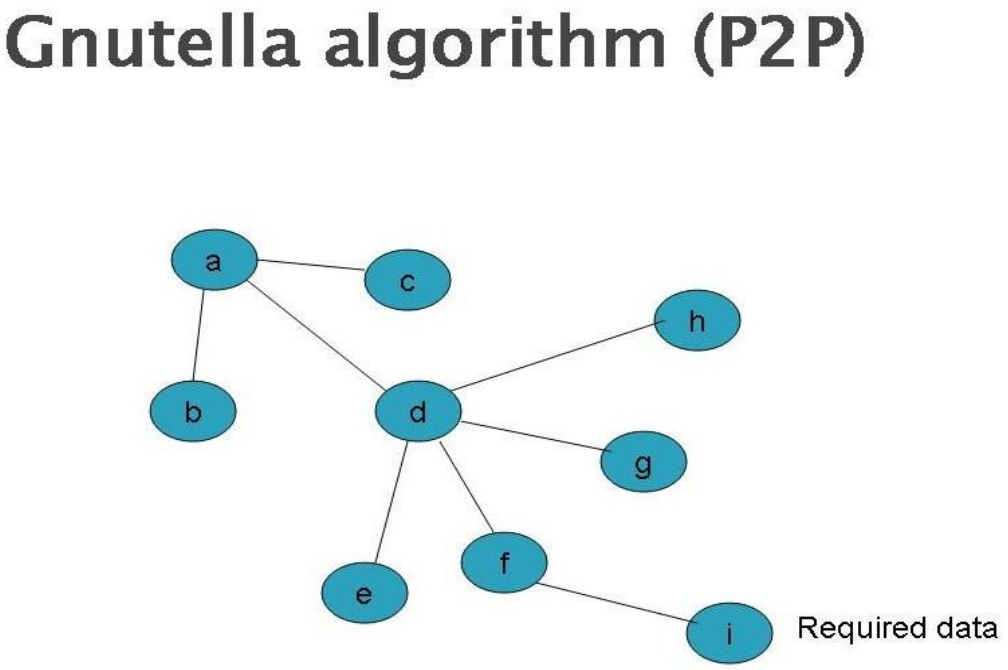


Figure-2 Illustration of a Peer to peer based routing architecture

V. Query routing methods

A. Forward knowledge systems

Effective query routing has to reduce both false positive and false negative results. In the context of information retrieval, false positive results are the non-relevant search results that are returned to the user. Such results will not help in fulfilling the user's information need. False negatives are those results that are very much relevant to the user's query but fail to be fetched by the system. Identifying false positives and false negatives an be done as a two fold process – one focusing on reducing false positive and another on minimizing false negative results.

a. Query refinement:

This is the process of specifying and modifying the query definition to focus on the information need. In some cases, the user is overwhelmed with unwanted information which is non-relevant to the query. This is possible because of broadly defined queries from a naïve user. Query refinement aims at refining and suggesting queries to the user in order to focus it only on the documents of interest. There are some approaches in refining the user queries:

i) Collocation of terms

This is the common way of refining the queries. This approach is a two-fold process where the set of documents containing terms from a user's query are identified and then query

suggestions are made with highest cumulative frequency of these documents. For example, For instance, in response to a query ``book suppliers", the query routing system will derive the following recommended terms: ``book store", ``book club", ``publisher".

ii) User Profiles

Another query refinement approach is to maintain profiles for the user using which the suggestions are derived for well focused queries. The main goal of this approach is to suggest replacement terms based on the semantic context and the scope of what the user requires from a particular query. These suggested terms are derived from domain-specific ontology or the user's feedback on the query context. Hence the recommendation is primarily based on the domain knowledge of the keywords present in the original user query. This approach reduces false positives and hence improves the efficiency and accuracy of the query refinement process.

The construction of the user query profiles can be done by using some domain-specific ontology or use an interactive dialog where the user can enter the synonyms of the query. This shows that the user profiles are not dependent on the information providers. So, even if the number of digital reference services increase, the answers are fetched seamlessly, thus enabling scalability.

b. Source Selection

This helps the users to identify and locate the correct digital reference service for a particular query thereby reducing the false negatives in the result. This mechanism aims at minimizing the overhead of reaching those digital reference services that do not satisfy the user query by pruning the non-relevant ones. Usually the user is interested in searching all the digital reference services that serve to satisfy the results.

For a given query, there can be many ways of searching for a particular information source. One way is to fetch all documents matching the query while other is to fetch those data sources which have high relevance – more matching documents than others. Given a query and search semantics, there are a number of approaches to find the best data sources for a particular query. This depends on the types and capabilities of the information sources. This is in alternate to searching all available information sources that contain matching documents.

i) Content summary:

This approach is popular in mediator based systems. Each router contains the summary of the terms which are handled by the data source associated with it. This enables the router to check for the relevance of the query with the content summary and thus redirecting it.

ii) Query Capability:

The query capabilities contain the capability descriptions of the source which are used to evaluate the relevance of a particular query. The data sources which contain the capability to handle the query and returning non-empty results are then chosen as relevant information sources for executing the query.

There are two main drawbacks in using both content summary and query capability. Systems using these methods require manual identification and collection of information on content summary and query capability. This is not feasible for large systems involving huge amount of information in a particular data source. The present systems which support capability based support do not perform query refinement which makes it ineffective to yield best routing results.

iii) Source Quality Measurements:

In order to select a particular information source, its quality measurements can be used. The quality dimensions include accuracy, reputation, timeliness, completeness, understandability of the sources. One disadvantage of this approach is its subjective nature. The quality of source selection will drop dramatically when the query users do not share the given set of information quality measurements.

The overall process of forward knowledge based systems and its work flow is shown in figure-3.

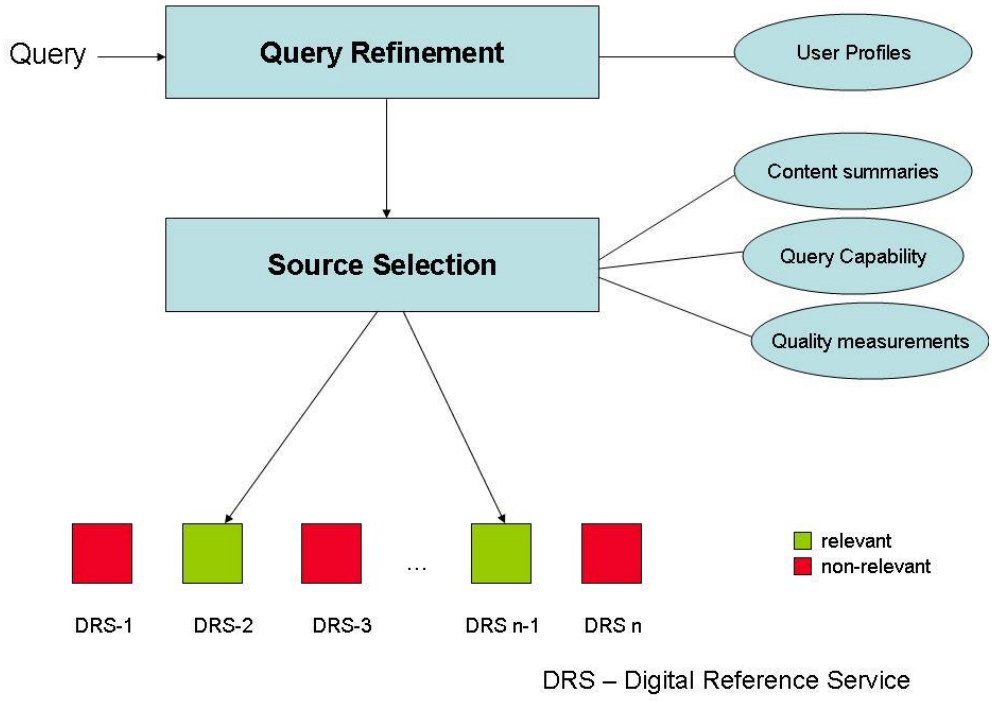


Figure-3 Work flow of Forward knowledge systems

B. Peer-to-Peer based routing

Peer to peer systems are the most popular medium through which huge amounts of data are shared. Their ability to build a resource-rich system by aggregating resources gives them more advantage than the centralized systems. The query forwarding problem of the distributed search system and the scalability limitations of the centralized search engines are overcome

by the peer based web search. Routing on these networks is either centralized or statically configured and is therefore unproblematic.

In the peer-based web search, each peer uses the results of its interaction with its neighbors to learn and rank them every time a query is encountered. This ranking is based on how well the other node in the network has returned the search results for a particular query. This model is used to dynamically route the queries according to the predicted match with other peer's knowledge. Also, the network topology is modified on the fly based on learned contexts and current information needs.

Overlay networks are a class of peer to peer networks where a virtual topology is built on top of physical links of the network. The uptime (active time) of a peer is relatively low. So, they leave and join the network whenever needed. This makes the topology of the network more dynamic. Routing in such systems is really problematic since a particular route will be valid only for a short period of time before a node in that route leaves the network.

a. Challenges in peer-to-peer routing algorithms:

Scalability: This is a measure of how a system performs when the number of nodes and hence the number of communications on the network grows.

Complexity: This is the measure of the number of intermediate steps involved before a packet travels from one peer node to another in a worst case scenario.

Anonymity: The anonymity factor is not applicable for all the peer based systems. There are certain systems which require the feature of anonymity. Handling the anonymity of peers should be done at the routing level.

b. Example Peer based models:

i) Gnutella Routing algorithm:

Gnutella is a file sharing network which uses overlay network model. The basic idea is that each node maintains a connection to a number of other nodes. In order to search any information in a network, the client broadcasts the message to all its peers. The node receiving the query, forwards it to its peers. This goes on until the required resource is found. The found resource is then returned back along the path sent. Since this process is recursive, the number of nodes queried has to be controlled. This is done using Time-to-Live counter. This implies the number of nodes or the time within which the resource has to be found. If the resource is not found within this limit, then a failure message is sent to client (where the query originated). This type of routing is the simplest kind possible for an overlay network. The network topology of Gnutella algorithm can be seen in figure-2.

This algorithm and network routing has many problems. The flooding of associated peers with query will work well for small or medium sized networks. But the cost of searching on this type of peer-to-peer network is more and as the network scales large, the cost increases exponentially. Gnutella was not designed to be anonymous and discovering who is making specific resources available is a simple matter of performing a search.

ii) Distributed Hash Tables:

Distributed Hash Table algorithms are decentralized distributed systems that use hash function for sharing key across peers. Any participating node can efficiently retrieve the value associated with a given key. DHT algorithms are useful for sharing files or other data across a peer-to-peer network. The hash function used here takes a string of variable length and returns a number that it generates. This algorithm works by hashing all data identifiers and storing their locations in a hash table. This table is distributed across all participating nodes.

Many DHTs use the flexibility to pick neighbors which are close in terms of latency in the physical underlying network. So, the maximum route length is closely related to the diameter, which is the maximum number of hops in any shortest path between nodes. Similar to graph theory, this algorithm is limited by the degree/diameter tradeoff. One main advantage of this algorithm is its time complexity of O (log n). It has lack of significant overhead which makes it better than the flooding algorithm. One good example of a DHT algorithm is chord, which is developed by a research group in MIT.

iii) Key-based routing (KBR)

Key-based routing is a lookup method which is used with overlay networks such as DHTs. DHTs provide a method to identify the host where the data is present. Key-based routing provides a method to find the closest host for the data which is present in host. The host is said to be closest by the number of network hops rather than the physical distance.

Freenet is one example of a distributed peer to peer network which is based on key-based routing. In this case, each file or data is associated with a key and files with similar keys are likely to cluster on similar set of nodes. The clusters are identified and the query is routed to them without needing to search many peers, thus minimizing the search space. This algorithm provides anonymity and efficient routing of information.

Each node on the network contains data which is not known by its peers. So, each node acts as a data store to which other nodes can read or write. This enables the network to act as a distributed file-system. The data on each node is encrypted and is not known to other nodes providing anonymity.

Each node has a routing table which has the address of its neighbors and their performance in returning particular keys. The performance is assessed by its response time and transfer time. When a node receives a request, it searches its routing table to find the node which has been most successful in returning a similar key before. It is the keys which are used for finding relevance and not the files.

There are many advantages in this type of routing, such as anonymity, improved efficiency over time etc. But the drawback is that the network is slow as the data must pass through all the intermediate nodes.

The query routing flow in a Freenet algorithm is discussed in figure-4.

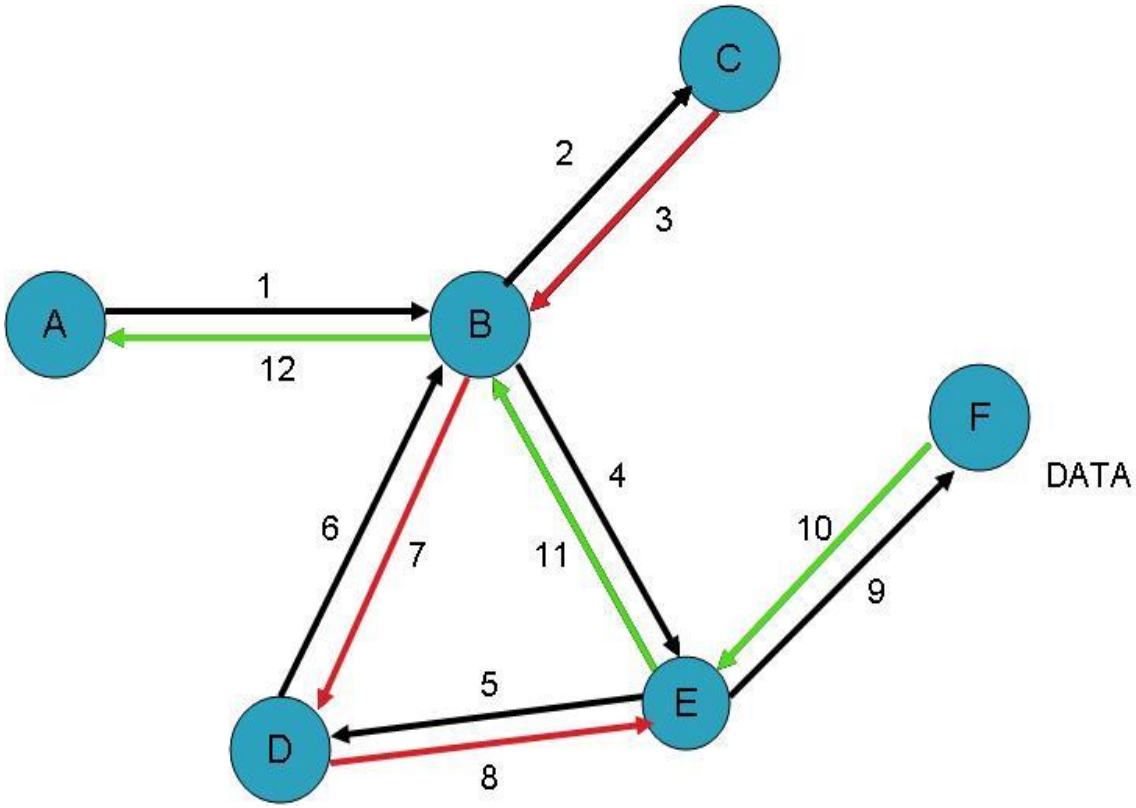


Figure-4 Query flow in Freenet routing algorithm

iv) Semantic Routing

Semantic routing is a method which is more focused on the nature of query to be routed and it does not count on the network topology. This algorithm prioritizes the nodes which have been previously good at providing information about the types of content referred to by the query.

In order to search for information in a peer to peer network using semantic routing, the data needs to have some semantic description associated with it. This usually is the metadata information for the data. Tagging the data improves the semantic nature in a peer-to-peer network.

This algorithm modifies the confidence on a particular peer if it answers a query. If the node has answered the query correctly, then the confidence measure is increased. There must be a constant identifier in the network's namespace to retain the confidence ratings. This is done in order to achieve persistence.

10. Resources

Required Readings:

For students:

Le-Shin Wu, Ruj Akavipat, Filippo Menczer. 2005. Adaptive Query Routing in Peer Web Search. *ACM* , 1074-1075. DOI= http://doi.acm.org/10.1145/1062745.1062875.

Arturo Crespo and Hector Garcia-Molina. 2002. Routing Indices For Peer-to-Peer Systems. Proceedings of the 22 nd International Conference on Distributed Computing Systems (ICDCS'02). IEEE Press. 23. DOI = http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.123.8939

Miguel Castro Peter, Peter Druschel, Ayalvadi Ganesh, Antony Rowstron, Dan S. Wallach. 2002. Secure routing for structured peer-to-peer overlay networks. 299-314. ACM. DOI = http://doi.acm.org/10.1145/1060289.1060317

Ling Liu. 1999. Query Routing in Large-scale Digital Library Systems. Computer Society. International Conference on Data Engineering (ICDE'99). IEEE Press. 154. DOI= http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.10.6346

Wai Gen Yee, Linh Thai Nguyen, Dongmei Jia, Ophir Frieder. 2008. Efficient query routing by improved peer description in P2P networks. International conference on Scalable information systems. (ICST'08). DOI= http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.143.4994

For Faculty:

A. Sugiura and O. Etzioni. 2000. Query Routing for Web Search Engines: Architecture and Experiments. North-Holland Publishing Co. 417 – 429. DOI = http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.79.522

Anjali Gupta, Barbara Liskov, Rodrigo Rodrigues. 2004. Efficient Routing for Peer-to-Peer Overlays. Proceedings of the 1st conference on Symposium on Networked Systems Design and Implementation - Volume 1. USENIX Association. 9. DOI = http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.1.5939

Gibbins, N. and Hall, W. 2001. Scalability Issues for Query Routing Service Discovery. Proceedings of the Second Workshop on Infrastructure for Agents, MAS and Scalable MAS. 209-217. DOI= http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.28.7362

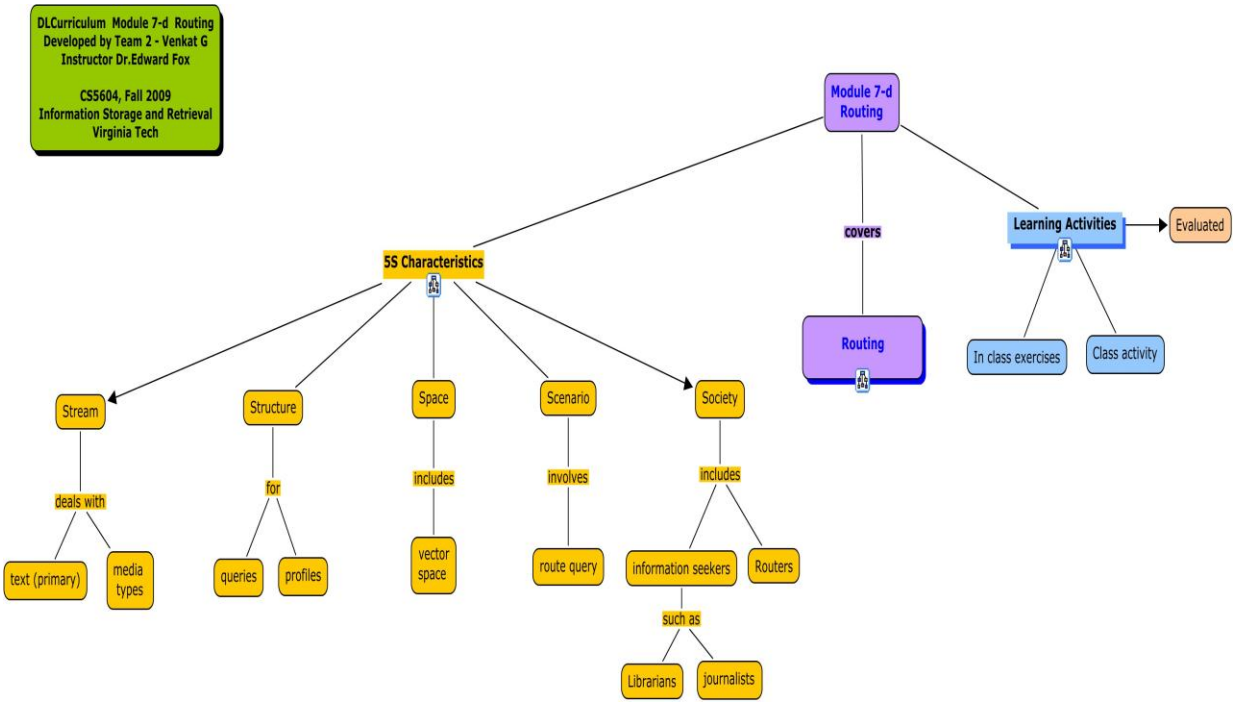
Abhishek Kumar, Jun (Jim) Xu, Ellen W. Zegura. 2005. Efficient and scalable query routing for unstructured peer-to-peer networks. 24th Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings IEEE. 1162- 1173. DOI= 10.1109/INFCOM.2005.1498343

Additional Useful Links:

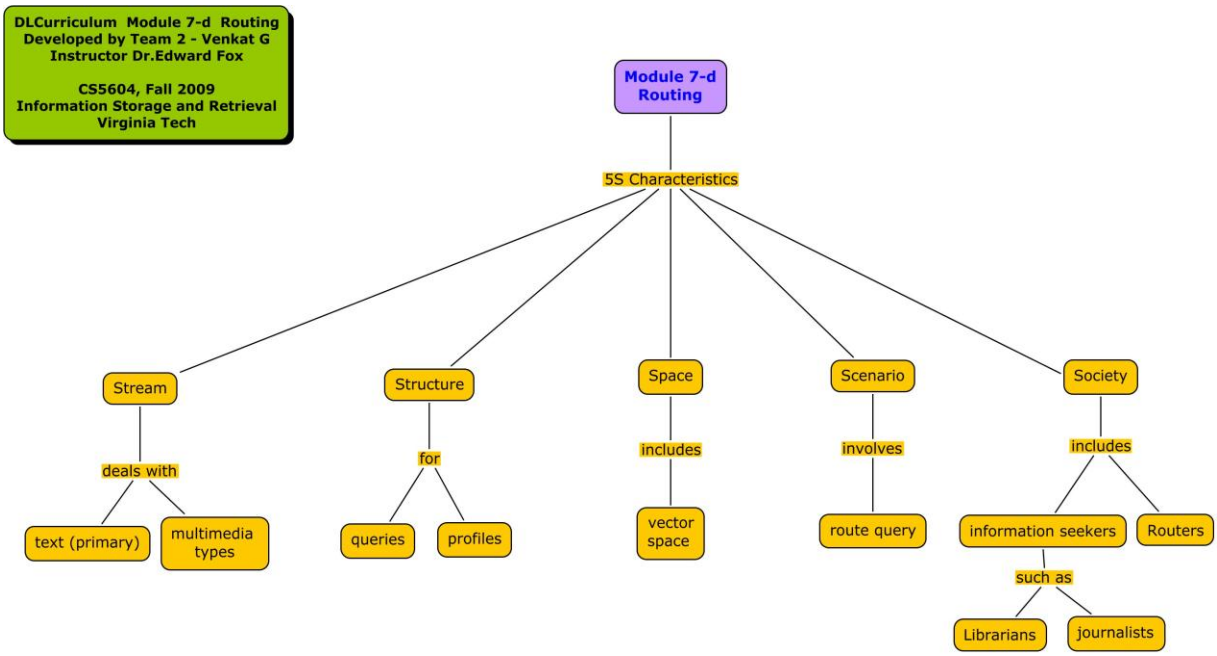
- Adaptive Query Routing at http://www.cc.gatech.edu/projects/disl/QR/
- Freenet at http://freenetproject.org/
- Chord at http://pdos.csail.mit.edu/chord/
- Peer to peer routing at http://ntrg.cs.tcd.ie/undergrad/4ba2.05/group6/index.html

11. Concept map

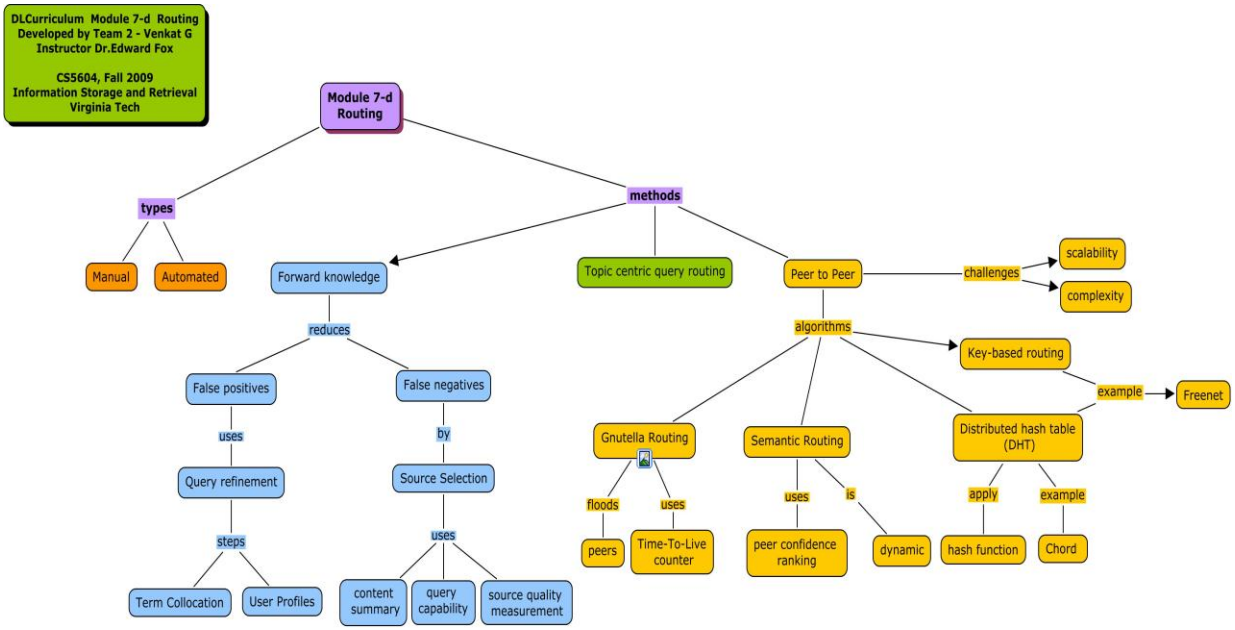
Overall Concept Map:



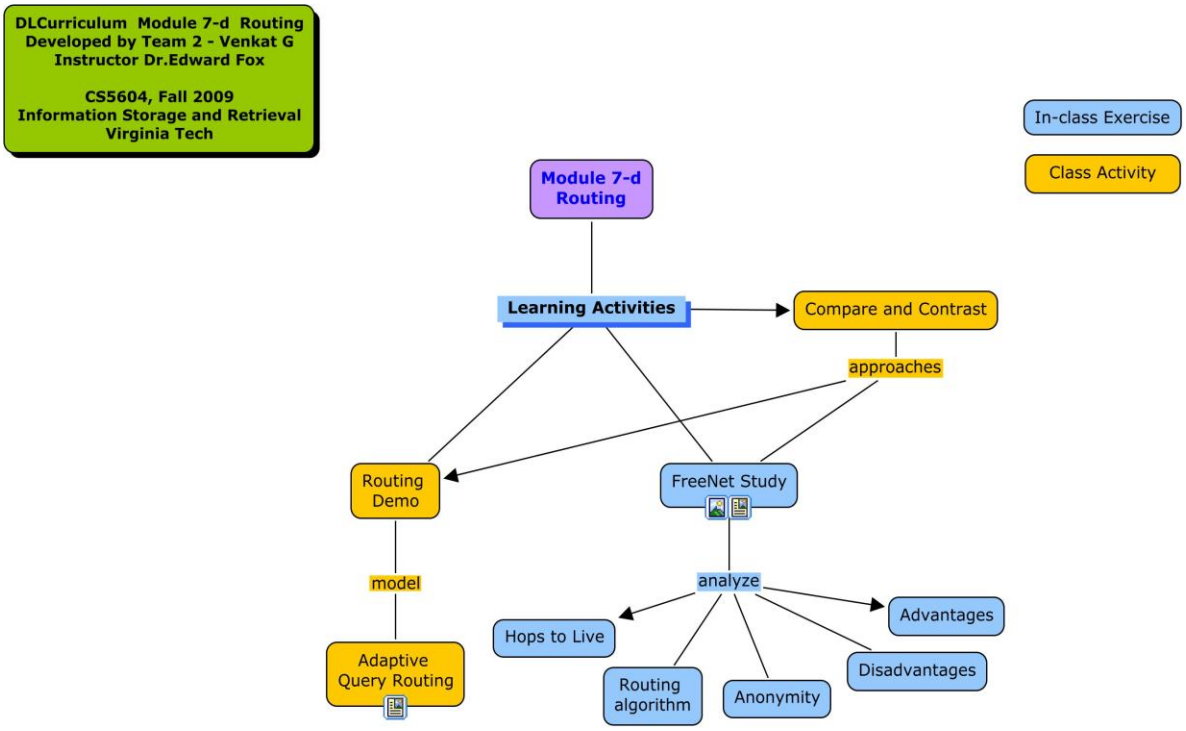
5S Characteristics:



Routing Types and Methods:



Learning Activities:



12. Exercises/Learning activities

a) Class Activity - Adaptive Query Routing demonstration (20 minutes)

Visit the link below to view the online demonstration on Adaptive Query Routing:
http://www.cc.gatech.edu/projects/disl/QR/

This is the homepage of the query routing software. AQR is an adaptive middleware layer that uses several query routing mechanisms and provides architecture for deploying them in an open networked environment. This online routing system was developed by Associate Professors Ling Liu and Carlton Pu with the help of graduate students David Buttler, Wei Han, Henrique Paques, and Wei Tang.

Select the 'walkthrough' section on the left panel. It navigates the user to the demo of query routing system. It provides an interactive demo of query routing aiding to understand the basic concepts and process involved in it. Break the class into 3-4 groups.

i) See the step-by-step process of query routing shown in walkthrough. Each group has to analyze and relate it to the techniques of query routing discussed in the syllabus 9.V.

(Hint: This system uses the process of query refinement and source selection.)

ii) Each group prepares a report on what happens in each stage of the routing process.

b) In-class Exercise - Analysis of Freenet (20 minutes)

Visit http://freenetproject.org/ which contains the Freenet application. This is based on peer to peer routing model.

Read the routing technique used in Freenet. Download the software and try it (This is optional because the software might consume a lot of disk space).

i) Identify the need for using routing in this application.

ii) What type of routing technique does this system employ?

iii) What is the purpose of using key-based algorithm in this application?

iv) How is anonymity achieved in this peer to peer model?

v) Identify the advantages and drawbacks of this model.

Prepare a report on the above questions.

More details on Freenet routing can be found in http://freenetproject.org/ngrouting.html

c) Class Activity – Compare and Contrast (15 minutes)

Compare and contrast the approaches used in 12 (a) and 12 (b). 12 (a) is a class activity on Adaptive Query Routing demonstration and 12 (b) is an exercise about Freenet. Both these systems follow different methodologies.

In class, break the students into groups of 3 or 4.

Have them discuss the differences in the methodologies used by AQR and Freenet.

Identify the advantages and disadvantages of the methods.

Let each group prepare a Word document or PowerPoint slides (5 minutes) and presents it to the class (10 minutes).

13. Evaluation of learning achievement

In their answers to the discussion questions, the students demonstrate an understanding of

- The need for routing systems
- The basic concepts of routing

- Different types and methodologies of routing techniques
- Challenges involved in routing queries

14. Glossary

- DRS: Digital Reference Service

- Clustering: grouping of documents which satisfy a set of common properties. The aim is to assemble together documents which are related among them. Clustering can be used, for instance, to expand a user query with new and related index terms.

- Scalability: The ability and ease with which a service may increase in size; where size may be defined as number of users, throughput of questions and responses, etc.

- Vector model: an algebraic model for representing text documents (and any objects, in general) as vectors of identifiers. It is used in information filtering, information retrieval, indexing and relevancy rankings.

15. Additional useful links

16. Contributors

Authors: Venkatasubramaniam Ganesan
Dr. Edward Fox

Evaluators: Ashwin Palani
Ashwin Khandeparker
Seungwon Yang
John Ewers