**Concept for a more advanced Search Engine:::CLUSTERS**

C.D. Jones

1. Use semantics in search so the meaning of the search is taken into account, not just raw parsing. Focus on understanding the search query, not just mechanics and mathematics.
2. Use natural language processing so it can do searching using **rephrased** versions of your keywords generating more, accurate, hits.
3. Often times when trying to search, ninety percent of the search process is determining what sort of query you should use. This shouldn’t be. Suggestions?
4. Layers. Multiple nested searches to narrow out what you are looking for. Example: First query is “Storms”, a normal results page comes up. Clicking on one of the results asks you if you want to keep that link or if you want to look further and search a nested layer of that first search. Then you get results, you can choose a link to go to or you can go further nested. This continues ad infinitum.
5. Queries are in the form of questions. The search engine is like talking to a computer like it is human. NLP is very important.
6. Use Clusters back end described below.
7. Real-time Search
8. Suggestions tailored by crossing previous searches and present search
9. Divides searches into subject matter. (Like Google has “Images”, “News”, “Maps”
10. To do this the data mining algorithm has to be able to classify pages and sift through subject matter.
11. The user is not “searching” but asking the search engine questions to which the search engine answers via the pages it returns.

Properties of Cluster Algorithm:

Common tags between two pages cause them to move together or “cluster”

Distance is a function of similar content.

Clusters are formed when individual pages have common terms. Entire clusters can be attracted to other clusters.

Large topics consist of major topics, while smaller clusters are subtopics.

Cluster space is defined using antonyms, where for example “hot” is on the opposite end of “Cold”. Antonym tags cause pages and clusters to repel.

**Example Clusters**

1.

<home>

Hi My Name is Ted and I’m building a search engine.

</home>

2.

<body>

Where is Joel? He’s late for his dentist appointment.

</body>

3.

<body>

I’m running a search on Joel in the search engine

</body>

Since the majority of movement is towards the origin, sending newly mined data to a pre-defined location representing the outskirts of the cluster space seems to make the most sense. The value for the outskirts boundary should be calculated empirically.

“It” Cluster

“Change” Cluster

Health Cluster

This data space can be useful for a different, more language based search engine.

Data Filter

Send to Cluster Space

Mine Data

Cluster Space

Search Engine Front End

**Data Mining**

*To do this the data mining algorithm has to be able to classify pages and sift through subject matter.*

What happens when the data miner is passed a new page? What does it do?

Classifies it. How?

When it data mines, first off, it pulls in all of the text on the page. The data miner gives the page a unique identifier. It then **calculates the meaning of the page**. This is the most difficult part. It finds the actors and actions in the sentences. Once it has accomplished the overall subject of the page, it tags it with various general annotations and sticks it in a cluster.

Once in a cluster?

For every data mined page, it has a general subject. Once entered into the cluster, it is given a three dimensional figure representing its location in cluster space. The cluster space updates all of the clusters according to the new addition to the space.

How do we handle queries?

The subject of the query must be determined. A **matching** process must occur where the subject is placed to a cluster that is closest in subject matter to that cluster. The clusters surrounding the query subject form the search results.

**Data Structures**

***Clusters***

Each cluster is a table entry in the clusters\_tbl. It has 3-tuple locating it in the cluster space. It has a text field containing its subject matter. It has a text field containing its URL.

There is a table containing the index of all the clusters. The table is search\_space\_tbl

After a tick, each cluster is examined in reference to its surrounding clusters and either moves towards one or away from another.