Alexandria Persing

Information Retrieval System

Experiments

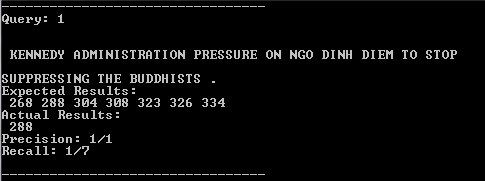
Definitions:

* Start/Base = the comparisons with a standard base tree, as first prototyped/designed
* COMPARE\_VALUE = the value at which the differences are so great that the trees are considered to be too different to continue
* BRANCH\_WEIGHT\_VALUE = the value which each branch’s weight is divided by before deciding how different the trees are; a higher value means that the branches are considered less than those which they stem from

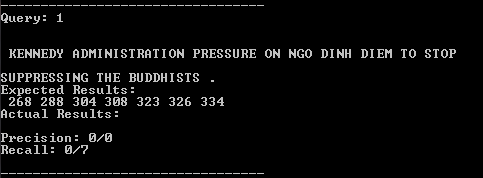
If not indicated, the following standard values were used for the test:  
COMPARE\_VALUE = .9;  
BRANCH\_WEIGHT\_VALUE = 2.7

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| What Changed from base / Hypothesis | Average Precision | Average Recall | Worst Precision | Worst Recall | Best Precision | Best Recall |
| Start/Base | .091 | .095 | 0 | 0 | 1 | 1 |
| Word Stemming | .067 | .085 | 0 | 0 | 1 | 1 |
| Multiple Word Search  (Boyer Moore string search) | .055 | .076 | 0 | 0 | 1 | 1 |
| Word Stemming + Multiple Word Search | .043 | .073 | 0 | 0 | 1 | 1 |
| Word Stemming + Multiple Word Search  COMPARE\_VALUE = 1;  BRANCH\_WEIGHT\_VALUE = 2.3; | .057 | .115 | 0 | 0 | 1 | 1 |
| Start/Base  COMPARE\_VALUE = 1;  BRANCH\_WEIGHT\_VALUE = 2.3; | .080 | .114 | 0 | 0 | 1 | 1 |
| Changed tree ontology - v2  COMPARE\_VALUE = 1;  BRANCH\_WEIGHT\_VALUE = 2.3; | .039 | .234 | 0 | 0 | .429 | 1 |
| Changed tree ontology - v2 | .038 | .210 | 0 | 0 | .375 | 1 |
| Changed tree ontology - v2  Word Stemming | .037 | .156 | 0 | 0 | .333 | 1 |
| Changed tree ontology - v2  Multiple Word Search | .050 | .380 | 0 | 0 | 1 | 1 |
| Changed tree ontology - v2  Word Stemming + Multiple Word Search | .040 | .397 | 0 | 0 | .667 | 1 |
| Changed tree ontology - v2  Word Stemming + Multiple Word Search  COMPARE\_VALUE = 1;  BRANCH\_WEIGHT\_VALUE = 2.3; | .042 | .429 | 0 | 0 | .667 | 1 |
| Changed tree ontology - v2  Word Stemming + Multiple Word Search  COMPARE\_VALUE = .8;  BRANCH\_WEIGHT\_VALUE = 2; | .019 | .054 | 0 | 0 | .571 | 1 |
| What Changed from base / Hypothesis | Average Precision | Average Recall | Worst Precision | Worst Recall | Best Precision | Best Recall |
| Changed tree ontology - v2  Word Stemming + Multiple Word Search  COMPARE\_VALUE = .8;  BRANCH\_WEIGHT\_VALUE = 1.5; | 0 | 0 | 0 | 0 | 0 | 0 |
| Changed tree ontology - v2  Word Stemming + Multiple Word Search  COMPARE\_VALUE = .8;  BRANCH\_WEIGHT\_VALUE = 1.9; | .021 | .041 | 0 | 0 | .8 | 1 |
| Changed tree ontology - v2  Word Stemming + Multiple Word Search  COMPARE\_VALUE = .8;  BRANCH\_WEIGHT\_VALUE = 1.8; | .020 | .015 | 0 | 0 | 1 | .6 |
| Changed tree ontology - v2  Word Stemming + Multiple Word Search  COMPARE\_VALUE = .9;  BRANCH\_WEIGHT\_VALUE = 1.9; | .031 | .090 | 0 | 0 | 1 | 1 |
| Changed tree ontology - v2  Word Stemming + Multiple Word Search  COMPARE\_VALUE = .9;  BRANCH\_WEIGHT\_VALUE = 1.8; | .017 | .066 | 0 | 0 | .444 | 1 |
| Changed tree ontology - v2  Word Stemming + Multiple Word Search  COMPARE\_VALUE = 1;  BRANCH\_WEIGHT\_VALUE = 1.8; | .059 | .175 | 0 | 0 | 1 | 1 |
| Changed tree ontology - v2  Word Stemming + Multiple Word Search  COMPARE\_VALUE = 1;  BRANCH\_WEIGHT\_VALUE = 1.9; | .079 | .241 | 0 | 0 | 1 | 1 |
| Changed tree ontology - v2  Word Stemming + Multiple Word Search  COMPARE\_VALUE = 1;  BRANCH\_WEIGHT\_VALUE = 2; | .085 | .290 | 0 | 0 | 1 | 1 |
| Changed tree ontology - v2  Word Stemming + Multiple Word Search  COMPARE\_VALUE = 1;  BRANCH\_WEIGHT\_VALUE = 2.1; | .075 | .333 | 0 | 0 | 1 | 1 |
| Tree ontology - v2  Word Stemming + Multiple Word Search (stemming multi words)  COMPARE\_VALUE = 1;  BRANCH\_WEIGHT\_VALUE = 2.1; | .051 | .308 | 0 | 0 | 1 | 1 |
| Tree ontology - v2  Word Stemming + Multiple Word Search (stemming multi words)  COMPARE\_VALUE = 1;  BRANCH\_WEIGHT\_VALUE = 2;  IRREL\_WEIGHT\_VALUE=2; | .032 | .484 | 0 | 0 | 1 | 1 |
| Tree ontology - v3  Word Stemming + Multiple Word Search (stemming multi words)  COMPARE\_VALUE = 1;  BRANCH\_WEIGHT\_VALUE = 2;  IRREL\_WEIGHT\_VALUE=2; | .044 | .449 | 0 | 0 | 1 | 1 |
| Tree ontology - v3  Word Stemming + Multiple Word Search (stemming multi words)  COMPARE\_VALUE = .9;  BRANCH\_WEIGHT\_VALUE = 2;  IRREL\_WEIGHT\_VALUE=2; | .042 | .333 | 0 | 0 | 1 | 1 |
| Tree ontology - v3  Word Stemming + Multiple Word Search (stemming multi words)  COMPARE\_VALUE = 1;  BRANCH\_WEIGHT\_VALUE = 2;  IRREL\_WEIGHT\_VALUE=1.5; | .030 | .372 | 0 | 0 | 1 | 1 |
|  |  |  |  |  |  |  |
| New Comparison tests: |  |  |  |  |  |  |
| V1,  COMPARE\_VALUE = .45;  BRANCH\_WEIGHT\_VALUE = 1.5; | .0745 | .313 | 0 | 0 | 1 | 1 |
| V1,  COMPARE\_VALUE = .5;  BRANCH\_WEIGHT\_VALUE = 1.2; | .059 | .194 | 0 | 0 | 1 | 1 |
| V1,  COMPARE\_VALUE = .4;  BRANCH\_WEIGHT\_VALUE = 1.6; | .061 | .296 | 0 | 0 | 1 | 1 |
| V1,  COMPARE\_VALUE = .45;  BRANCH\_WEIGHT\_VALUE = 1.4; | .0948 | .273 | 0 | 0 | 1 | 1 |
| V2,  COMPARE\_VALUE = .45;  BRANCH\_WEIGHT\_VALUE = 1.4;  ERROR\_LEVEL=.15; | .073 | .645 | 0 | 0 | 1 | 1 |
| V2,  COMPARE\_VALUE = .4;  BRANCH\_WEIGHT\_VALUE = 1.4;  ERROR\_LEVEL=.15; | .0931 | .607 | 0 | 0 | 1 | 1 |
| V2,  COMPARE\_VALUE = .4;  BRANCH\_WEIGHT\_VALUE = 1.4;  ERROR\_LEVEL=.14; | .101 | .584 | 0 | 0 | 1 | 1 |
| V2,  COMPARE\_VALUE = .35;  BRANCH\_WEIGHT\_VALUE = 1.4;  ERROR\_LEVEL=.13; | .129 | .465 | 0 | 0 | 1 | 1 |
| V2,  COMPARE\_VALUE = .35;  BRANCH\_WEIGHT\_VALUE = 1.3;  ERROR\_LEVEL=.13; | .131 | .393 | 0 | 0 | 1 | 1 |
| V2,  COMPARE\_VALUE = .37;  BRANCH\_WEIGHT\_VALUE = 1.4;  ERROR\_LEVEL=.14; | .108 | .553 | 0 | 0 | 1 | 1 |
| V2,  COMPARE\_VALUE = .37;  BRANCH\_WEIGHT\_VALUE = 1.35;  ERROR\_LEVEL=.14; | .122 | .527 | 0 | 0 | 1 | 1 |
| V3,  COMPARE\_VALUE = .37;  BRANCH\_WEIGHT\_VALUE = 1.35;  ERROR\_LEVEL=.14; | .138 | .462 | 0 | 0 | 1 | 1 |
| V3, tree w/ more words  COMPARE\_VALUE = .37;  BRANCH\_WEIGHT\_VALUE = 1.35;  ERROR\_LEVEL=.14; | .142 | .417 | 0 | 0 | 1 | 1 |
| V3, tree w/ more words  COMPARE\_VALUE = .35;  BRANCH\_WEIGHT\_VALUE = 1.35;  ERROR\_LEVEL=.14; | .139 | .389 | 0 | 0 | 1 | 1 |
| V3, tree w/ more words  COMPARE\_VALUE = .38;  BRANCH\_WEIGHT\_VALUE = 1.35;  ERROR\_LEVEL=.14; | .141 | .462 | 0 | 0 | 1 | 1 |
| V4, restructured tree  COMPARE\_VALUE = .37;  BRANCH\_WEIGHT\_VALUE = 1.35;  ERROR\_LEVEL=.14; | .099 | .311 | 0 | 0 | 1 | 1 |
| COMPARE\_VALUE = .33;  BRANCH\_WEIGHT\_VALUE = 1.2;  ERROR\_LEVEL=.14;  Not found weight: 1.1 | .189 | .14 | 0 | 0 | 1 | 1 |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |

Before (Tree V2): [Comparator values: 1, 2]

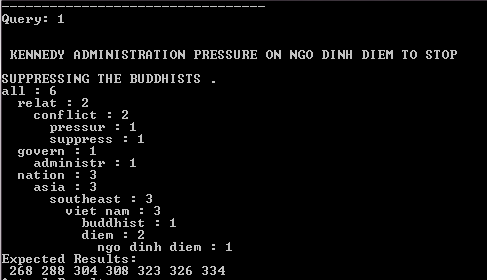


After adding more words specifically in the articles to the content tree:



Problem?

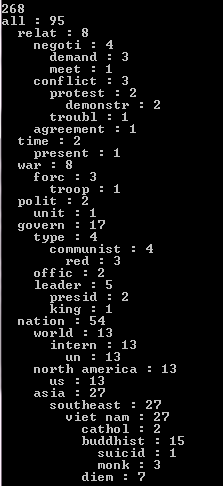
Matching queries & documents when queries have significantly fewer words in them



Relations: 1/3 ~ .333

Government: 1/6 ~ .1667

Nation: ½ ~.5



Relat: 8/95 ~ .084

Time: 2/95 ~ .021

War: 8/95 ~ .084

Polit: 2/95 ~ .021

Govern: 17/95 ~ .179

Nation: 54/95 ~ .568

In the first branches alone, the differences work out to where relations are already significantly different, such that there will be no matches

When you compare all of the documents against one another with the values necessary to get results from queries, each document has 100-200 results of other documents which match it.

In order to get a more reasonable representation (~10 matches each), you have to change the compare and branch weight values to .8 and 1.5 respectively, resulting in much closer matches.

At these levels, the queries get no results, because their weighting for each branch is comparatively much larger, given the fewer number of total mapped words.

Also, this implies that any information which the article contains which is extraneous compared to the query at hand is counted against it as becoming more irrelevant. Weighting branches which are irrelevant as less failed to improve precision – to be more accurate, it actually decreased precision.

In addition, many times there are words which imply a cause-effect relationship which is not able to be accurately represented within the tree.

Proof that using the documents themselves as queries works better: 

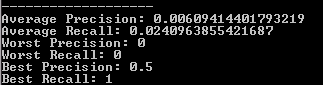
When each document is ran against all other documents (using values of .8 and 1.5), and we assume that documents which are both expected answers for the same query are probably relevant to one another, the average precision more than doubles from that of any attempts for queries.

Flat tree compare all:



Precision actually improves, but recall decreases

Flat tree queries:



As expected, precision significant decreases, as well as recall

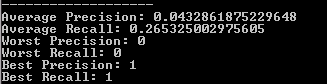
More structured tree:

Compare all



Precision is more than the previous structured tree, but still less than the flat tree  
Recall is better than either the previous tree or the flat tree

Queries



Precision is greater than the flat tree, as well as the recall

New algorithm:

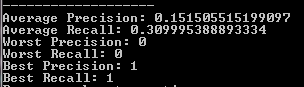
 compare all (.15 compare, 1.35 branch weight, .14 error level)

See above for query values

Precision for queries w/ recall > .5:

|  |  |
| --- | --- |
| V3, tree w/ more words  COMPARE\_VALUE = .38;  BRANCH\_WEIGHT\_VALUE = 1.35;  ERROR\_LEVEL=.14; | .07 |
| V3, tree w/ more words  COMPARE\_VALUE = .2;  BRANCH\_WEIGHT\_VALUE = 1.35;  ERROR\_LEVEL=.14; | .03 |
|  |  |
|  |  |
|  |  |
|  |  |

AutoTree @ .05 idf min, basic trees



AutoTree stemmed max : .06