UNIVERSITÀ DEGLI STUDI DI SALERNO
DIPARTIMENTO DI INFORMATICA

Master Course in Computer Science

Social Networks

Professors:
Vincenzo Auletta
Diodato Ferraioli

Team Members:
Amedeo Leo
Francesco Gaetano
Luigi Lomasto

Fisciano 30/06/2016

1 Best Match

The best match algorithm provides an useful approach to simplify the way a generic individual performs his research based on a given input.

Using the best match strategy allows us to efficiently find the documents that mostly fit to the query we give in input to the algorithm.

For example, assuming that we want to make a classification of the pages based on the top-15 topics, if our query input is *"Obama Says Climate Change Already Damaging National Parks"*, we expect that the most influential output pages will be inherent to the "news" topic. Whereas, if our input is a sentence like the following: *"Obama stays aboard Air Force One to final minutes of Game 7"*, is reasonable to assume that one or more pages related to the "game" topic could appear between the first output documents.

The dataset we have used as input for the tests was provided by the Wibbi online crawler. This gave us, for each top-15 topic, a set of visited pages by the crawler and the relative text contents as well. Afterwards, we used an ad-hoc parser built upon some regular expressions, in order to remove all the punctuation and undesired words, such as the stopwords and any CSS tags and keywords.

Furthermore, the parser produces in output the files properly structured which we have used as input for the algorithm, as we will see in the following sections. Finally, the parser works offline: it means that we runned it once ad we won't consider the time of parsing in our tests.

1.1 Basic best match

This is the basic version of the algorithm seen during the course. The algorithm takes in input the first one of the files produced by the parser, whose structure is the following:

doc_name_i word_1,word_2,...,word_n

where:

- *doc_i* is the url of the i-th page visited by the crawler, with $1 < i < N$, where N is the number of the documents
- *word_j* is the j-th word that occurs in the page, with $1 < j < n$, where n is the number of words for each document

For each document we compute a score depending on the frequency of a query term in that document, which is the ratio between the number of occurrences of the term and the total number of words in the document. Afterwards, the algorithm returns the 20 documents with higher scores.

1.2 Optimized best match

In order to reduce the commitment required from the computation of all documents, an optimization of the basic algorithm has been implemented.

The idea is based on the fact that we should consider only a subset of the whole documents. For this reason we now consider an *inverted index structure* which, for each word that occurs in the complete dataset, tracks the set of documents in which it is contained. The algorithm takes in input the second file of the parser whose structure is the following:

term_i doc_1,f_1;doc_2,f_2...,doc_N

where:

- *term_i* is the i-th word that occurs in the dataset, with 1 < i < n, where n is the total number of words in the dataset
- *doc_j* is the j-th document in which term is contained and *f_j* is the related frequency of occurrences computed offline by the parser, with 1 < j < N, where N is the number of pages analyzed

For each query term in input, the score is evaluated only for a fixed threshold K=0,2% of the documents. Afterwards, the algorithm returns the 20 documents with higher scores.

2 Page Rank

Page Rank is a type of algorithm used by modern search engine, which assigns a numerical weight to web pages in order to retrieve the page's relevance and importance.

The main idea is that the importance of a web page depends on the importance of all the pages having a link to it. The intent is that the higher the Page Rank of a page, the more relevant it is.

The parser described previously has been exploited to generate offline an input file properly structured for the page rank algorithm:

topic_i doc_1;neigh_1,...,neigh_n*doc_2*...*doc_N

where:

- *topic_i* is the i-th read topic, for $1< i < 15$
- *doc_j* is the j-th url of the document related to the current topic, for $1 < j < N$ where N is the total number of pages
- *neigh_l* is the l-th url of the neighbour of the current document, for $1 < l < n$ where n is the size of the neighbourhood of the current document

In order to create a more realistic network graph, we chose ten random pages for each topic, linking them to ten pages belonged to another topic.

2.1 Basic Page Rank

This is the version of the algorithm seen during the course. This solution comprises the **"random walk"** strategy in order to avoid eventually *spider traps* (a set of nodes with no arcs out), which cause a wrong calculation of the ranks.

The algorithm takes in input a value ß, with 0,80 < ß < 0,89.

The calculation of the rank is the following:

1. ***R(0)***: (1-ß)/n at step i=0
2. ***R(i+1)***: (ß*R(i))/n for each step i > 0

where n is the number of all pages in the graph.

The term ßR(i) represents the case where, with probability ß, the random walker decides to follow an out link from his present page.

Instead, the initialization $(1 - ß)/n$ represents the probability of starting a random walk (transportation to a random pages).

2.2 Topic Sensitive Page Rank

We implemented a different approach to the computation of the page's rank. The main idea is to make a distinction between pages belonged to different topics; this is a variation about the way random walk are performed. Basically, we prefer to land on a page that is known to cover the chosen topic. For this reason, the topic-sensitive approach creates a different rank vector for each topic, forcing a page to spread exclusively his rank to the other pages related to that topic.

According to this, we have done a partition of the documents for each different topic. Doing so, we will have a S_i set of documents for 1 < i < 15, which are called *teleport sets*.

The variation in computing the rank is the following:

1. ***R(0)***: (1-ß)/s_j at step i=0
2. ***R(i+1)***: (ß*R(i))/s_j for each step i > 0

Where s_j is the cardinality of S_j. Those operations are repeated for each topic. Finally, we had to consider the ten pages linking to another topic. For example, if a sport page links to a news page, we would add to the total rank of the latest also the portion of rank given by the first. That describes the (unlikely) case where we travel from a topic page to another.

3 Match/Rank testing

In this section we are going to discuss about the results we obtained running both the algorithms and relative optimizations seen above.

3.1 Testing approach

We implemented a class tester whose purpose is to make a combination between each of the matching and rank algorithms.

For this reason we have four different combine methods:

1. **C1**: The output of the basic best match is given as input to the basic page rank

2. **C2**: The output of the optimized best match is given as input to the basic page rank

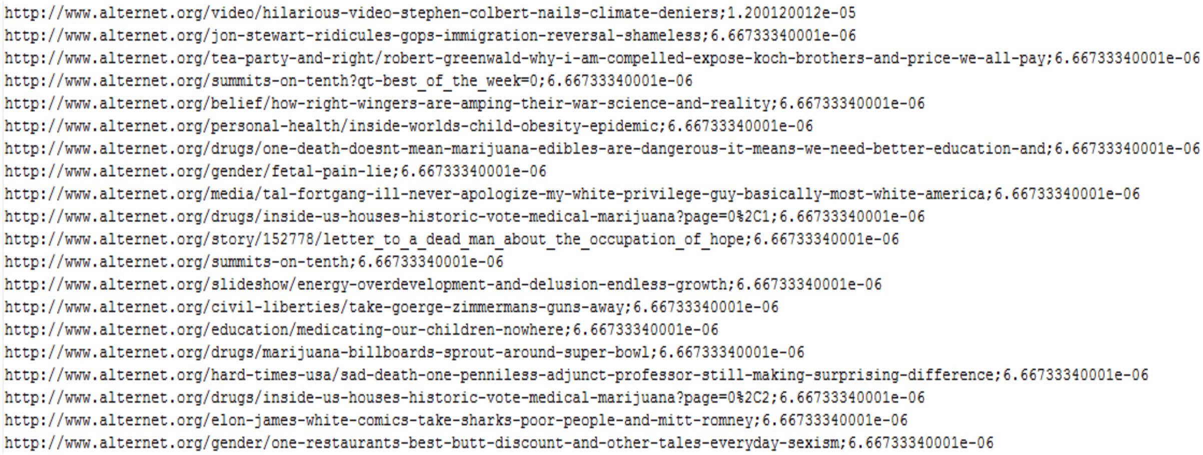
3. **C3**: The output of the basic best match is given as input to the topic sensitive page rank

4. **C4**: The output of the optimized best match is given as input to the topic sensitive page rank

The role of each combine is to extract the 20 documents given by the match algorithm from the graph outcome of the rank match and sort them in decreasing order. Furthermore, we will show the times and steps required from the algorithms to terminate, in order to understand which one is the most efficient.

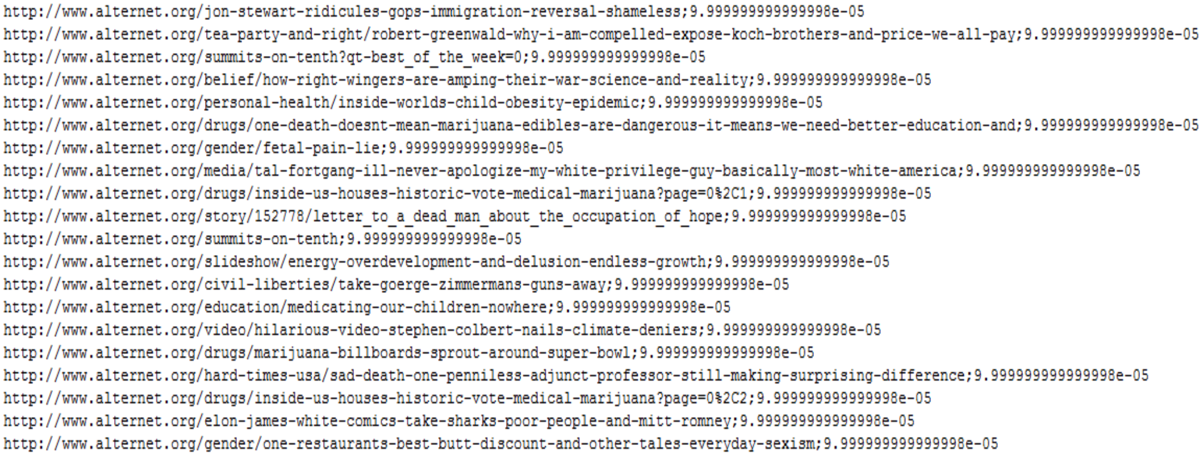
3.2 Results

Firstly, we will see how the algorithms behave running them on a simple query composed by a single word. For example, we want to submit the word ”yogurt”. Clearly, the output we expect will be composed mostly by pages inherent to the ”health” topic. The following is the output of both basic best match and page rank:



```
http://www.alternet.org/video/hilarious-video-stephen-colbert-nails-climate-deniers;1.200120012e-05
http://www.alternet.org/jon-stewart-ridicules-gops-immigration-reversal-shameless;6.66733340001e-06
http://www.alternet.org/tea-party-and-right/robert-greenwald-why-i-am-compelled-expose-koch-brothers-and-price-we-all-pay;6.66733340001e-06
http://www.alternet.org/summits-on-tenth?qt-best_of_the_week=0;6.66733340001e-06
http://www.alternet.org/belief/how-right-wingers-are-amping-their-war-science-and-reality;6.66733340001e-06
http://www.alternet.org/personal-health/inside-worlds-child-obesity-epidemic;6.66733340001e-06
http://www.alternet.org/drugs/one-death-doesnt-mean-marijuana-edibles-are-dangerous-it-means-we-need-better-education-and;6.66733340001e-06
http://www.alternet.org/gender/fetal-pain-lie;6.66733340001e-06
http://www.alternet.org/media/tal-fortgang-ill-never-apologize-my-white-privilege-guy-basically-most-white-america;6.66733340001e-06
http://www.alternet.org/drugs/inside-us-houses-historic-vote-medical-marijuana?page=0%2C1;6.66733340001e-06
http://www.alternet.org/story/152778/letter_to_a_dead_man_about_the_occupation_of_hope;6.66733340001e-06
http://www.alternet.org/summits-on-tenth;6.66733340001e-06
http://www.alternet.org/slideshow/energy-overdevelopment-and-delusion-endless-growth;6.66733340001e-06
http://www.alternet.org/civil-liberties/take-goerge-zimmermans-guns-away;6.66733340001e-06
http://www.alternet.org/education/medicating-our-children-nowhere;6.66733340001e-06
http://www.alternet.org/drugs/marijuana-billboards-sprout-around-super-bowl;6.66733340001e-06
http://www.alternet.org/hard-times-usa/sad-death-one-penniless-adjunct-professor-still-making-surprising-difference;6.66733340001e-06
http://www.alternet.org/drugs/inside-us-houses-historic-vote-medical-marijuana?page=0%2C2;6.66733340001e-06
http://www.alternet.org/elon-james-white-comics-take-sharks-poor-people-and-mitt-romney;6.66733340001e-06
http://www.alternet.org/gender/one-restaurants-best-butt-discount-and-other-tales-everyday-sexism;6.66733340001e-06
```

Figure 1: Output for basic best match / basic page rank

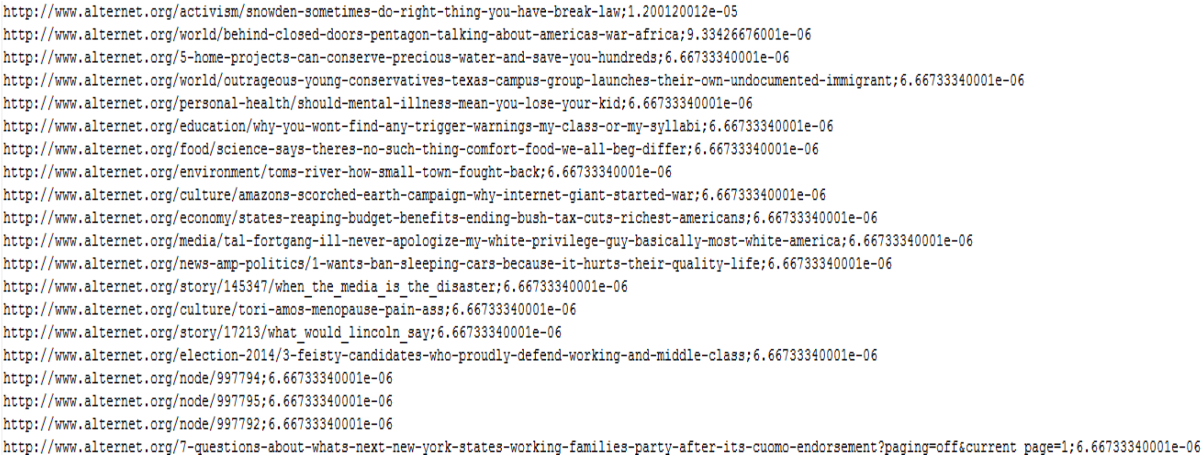


```
http://www.alternet.org/jon-stewart-ridicules-gops-immigration-reversal-shameless;9.999999999999998e-05
http://www.alternet.org/tea-party-and-right/robert-greenwald-why-i-am-compelled-expose-koch-brothers-and-price-we-all-pay;9.999999999999998e-05
http://www.alternet.org/summits-on-tenth?qt-best_of_the_week=0;9.999999999999998e-05
http://www.alternet.org/belief/how-right-wingers-are-amping-their-war-science-and-reality;9.999999999999998e-05
http://www.alternet.org/personal-health/inside-worlds-child-obesity-epidemic;9.999999999999998e-05
http://www.alternet.org/drugs/one-death-doesnt-mean-marijuana-edibles-are-dangerous-it-means-we-need-better-education-and;9.999999999999998e-05
http://www.alternet.org/gender/fetal-pain-lie;9.999999999999998e-05
http://www.alternet.org/media/tal-fortgang-ill-never-apologize-my-white-privilege-guy-basically-most-white-america;9.999999999999998e-05
http://www.alternet.org/drugs/inside-us-houses-historic-vote-medical-marijuana?page=0%2C1;9.999999999999998e-05
http://www.alternet.org/story/152778/letter_to_a_dead_man_about_the_occupation_of_hope;9.999999999999998e-05
http://www.alternet.org/summits-on-tenth;9.999999999999998e-05
http://www.alternet.org/slideshow/energy-overdevelopment-and-delusion-endless-growth;9.999999999999998e-05
http://www.alternet.org/civil-liberties/take-goerge-zimmermans-guns-away;9.999999999999998e-05
http://www.alternet.org/education/medicating-our-children-nowhere;9.999999999999998e-05
http://www.alternet.org/video/hilarious-video-stephen-colbert-nails-climate-deniers;9.999999999999998e-05
http://www.alternet.org/drugs/marijuana-billboards-sprout-around-super-bowl;9.999999999999998e-05
http://www.alternet.org/hard-times-usa/sad-death-one-penniless-adjunct-professor-still-making-surprising-difference;9.999999999999998e-05
http://www.alternet.org/drugs/inside-us-houses-historic-vote-medical-marijuana?page=0%2C2;9.999999999999998e-05
http://www.alternet.org/elon-james-white-comics-take-sharks-poor-people-and-mitt-romney;9.999999999999998e-05
http://www.alternet.org/gender/one-restaurants-best-butt-discount-and-other-tales-everyday-sexism;9.999999999999998e-05
```

Figure 2: Output for basic best match / topic sensitive page rank

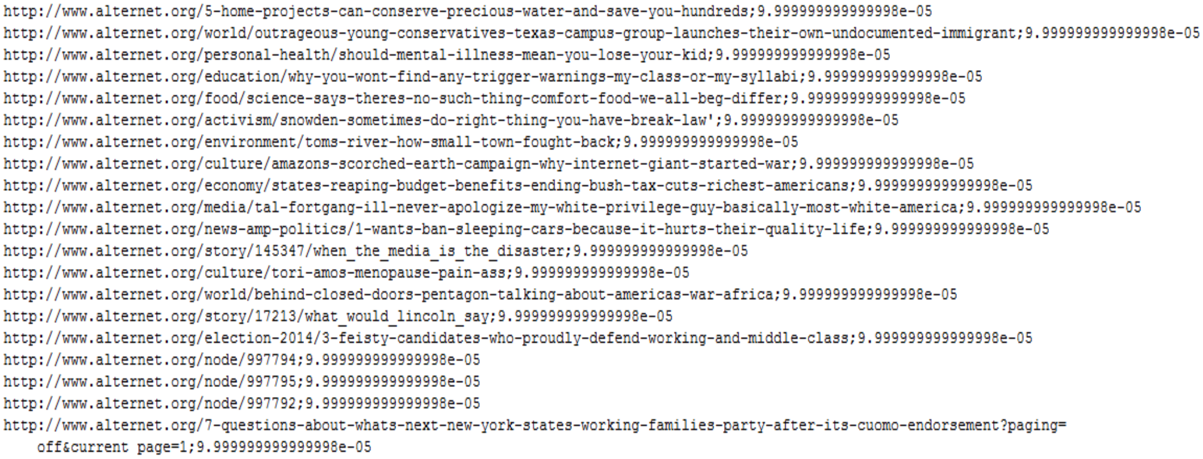
We can notice that the documents given are the same (because that is the output of the best match algorithm), although the order is different because the value of rank changes every time we use the two methods.

A similar result we can see if we run both the page rank methods on the optimized best match algorithm:



```
http://www.alternet.org/activism/snowden-sometimes-do-right-thing-you-have-break-law;1.200120012e-05
http://www.alternet.org/world/behind-closed-doors-pentagon-talking-about-americas-war-africa;9.33426676001e-06
http://www.alternet.org/5-home-projects-can-conserve-precious-water-and-save-you-hundreds;6.66733340001e-06
http://www.alternet.org/world/outrageous-young-conservatives-texas-campus-group-launches-their-own-undocumented-immigrant;6.66733340001e-06
http://www.alternet.org/personal-health/should-mental-illness-mean-you-lose-your-kid;6.66733340001e-06
http://www.alternet.org/education/why-you-wont-find-any-trigger-warnings-my-class-or-my-syllabi;6.66733340001e-06
http://www.alternet.org/food/science-says-theres-no-such-thing-comfort-food-we-all-beg-differ;6.66733340001e-06
http://www.alternet.org/environment/toms-river-how-small-town-fought-back;6.66733340001e-06
http://www.alternet.org/culture/amazons-scorched-earth-campaign-why-internet-giant-started-war;6.66733340001e-06
http://www.alternet.org/economy/states-reaping-budget-benefits-ending-bush-tax-cuts-richest-americans;6.66733340001e-06
http://www.alternet.org/media/tal-fortgang-ill-never-apologize-my-white-privilege-guy-basically-most-white-america;6.66733340001e-06
http://www.alternet.org/news-amp-politics/1-wants-ban-sleeping-cars-because-it-hurts-their-quality-life;6.66733340001e-06
http://www.alternet.org/story/145347/when_the_media_is_the_disaster;6.66733340001e-06
http://www.alternet.org/culture/tori-amos-menopause-pain-ass;6.66733340001e-06
http://www.alternet.org/story/17213/what_would_lincoln_say;6.66733340001e-06
http://www.alternet.org/election-2014/3-feisty-candidates-who-proudly-defend-working-and-middle-class;6.66733340001e-06
http://www.alternet.org/node/997794;6.66733340001e-06
http://www.alternet.org/node/997795;6.66733340001e-06
http://www.alternet.org/node/997792;6.66733340001e-06
http://www.alternet.org/7-questions-about-whats-next-new-york-states-working-families-party-after-its-cuomo-endorsement?paging=off&current page=1;6.66733340001e-06
```

Figure 3: Output of optimized best match / basic page rank



```
http://www.alternet.org/5-home-projects-can-conserve-precious-water-and-save-you-hundreds;9.999999999999998e-05
http://www.alternet.org/world/outrageous-young-conservatives-texas-campus-group-launches-their-own-undocumented-immigrant;9.999999999999998e-05
http://www.alternet.org/personal-health/should-mental-illness-mean-you-lose-your-kid;9.999999999999998e-05
http://www.alternet.org/education/why-you-wont-find-any-trigger-warnings-my-class-or-my-syllabi;9.999999999999998e-05
http://www.alternet.org/food/science-says-theres-no-such-thing-comfort-food-we-all-beg-differ;9.999999999999998e-05
http://www.alternet.org/activism/snowden-sometimes-do-right-thing-you-have-break-law';9.999999999999998e-05
http://www.alternet.org/environment/toms-river-how-small-town-fought-back;9.999999999999998e-05
http://www.alternet.org/culture/amazons-scorched-earth-campaign-why-internet-giant-started-war;9.999999999999998e-05
http://www.alternet.org/economy/states-reaping-budget-benefits-ending-bush-tax-cuts-richest-americans;9.999999999999998e-05
http://www.alternet.org/media/tal-fortgang-ill-never-apologize-my-white-privilege-guy-basically-most-white-america;9.999999999999998e-05
http://www.alternet.org/news-amp-politics/1-wants-ban-sleeping-cars-because-it-hurts-their-quality-life;9.999999999999998e-05
http://www.alternet.org/story/145347/when_the_media_is_the_disaster;9.999999999999998e-05
http://www.alternet.org/culture/tori-amos-menopause-pain-ass;9.999999999999998e-05
http://www.alternet.org/world/behind-closed-doors-pentagon-talking-about-americas-war-africa;9.999999999999998e-05
http://www.alternet.org/story/17213/what_would_lincoln_say;9.999999999999998e-05
http://www.alternet.org/election-2014/3-feisty-candidates-who-proudly-defend-working-and-middle-class;9.999999999999998e-05
http://www.alternet.org/node/997794;9.999999999999998e-05
http://www.alternet.org/node/997795;9.999999999999998e-05
http://www.alternet.org/node/997792;9.999999999999998e-05
http://www.alternet.org/7-questions-about-whats-next-new-york-states-working-families-party-after-its-cuomo-endorsement?paging=
    off&current page=1;9.999999999999998e-05
```

Figure 4: Output of optimized best match / topic sensitive page rank

The performance of the algorithms in terms of step and time elapsed are the following:

The x-axis indicates the values of ß (taxation parameter) for each test runned.

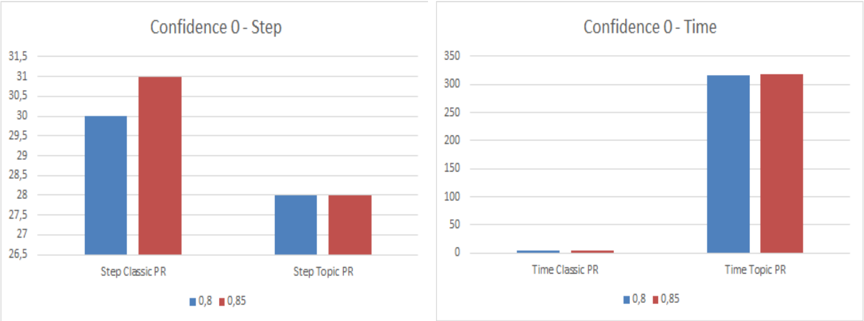


Figure 5: Steps and time elapsed for page rank and topic sensitive

As we can clearly see the graphics show two opposite outcomes. The basic page rank takes less time then the topic sensitive, even though it needs more steps to converge. It is normal if we consider how topic sensitive works. For each step it has to update a different rank vector for each topic, although it allows the algorithm to converge faster to the solution. We suppose that a parallel implementation of the topic sensitive algorithm might lead to more high performances.

Now we want to see for only two values of the taxation parameters the difference of time and steps to converge if we consider small convergence values (i.e. at step i, the algorithm stops if the total difference of values between rank of step i and rank of step i-1 is less or equal the confidence).

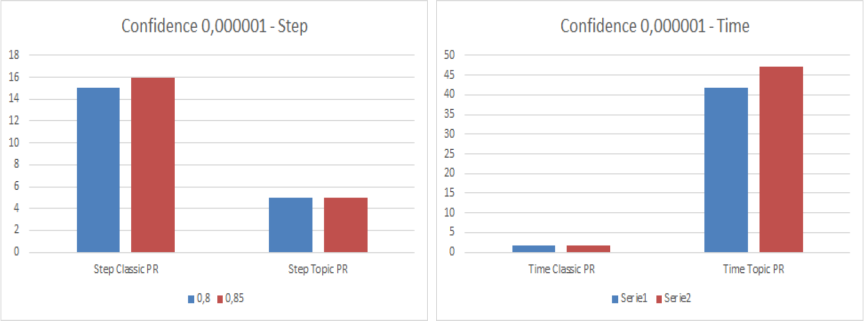


Figure 6: time/step for 0,000001 confidence value

We have a substantial reduction of steps required for both algorithms even for a very small confidence number. We think it depends on the fact that all rank values are very small and closer to each other. The consequence is that also the elapsed time results lower.

We tried to further augment the confidence:

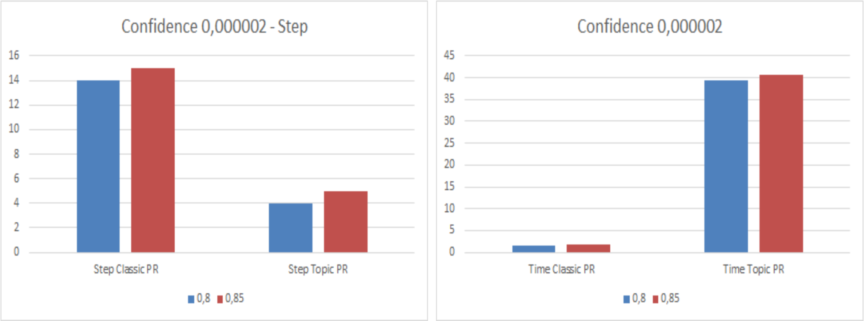


Figure 7: time/step for 0,000002 confidence value

As we can see, there is no significant difference from the last case, although even in this case we have a slight values' reduction.

The second test we want to consider also was runned on a one-term query: ”medicine”, even though this time is a more specific word. We tried this test to see if there are diversifications from the last test.

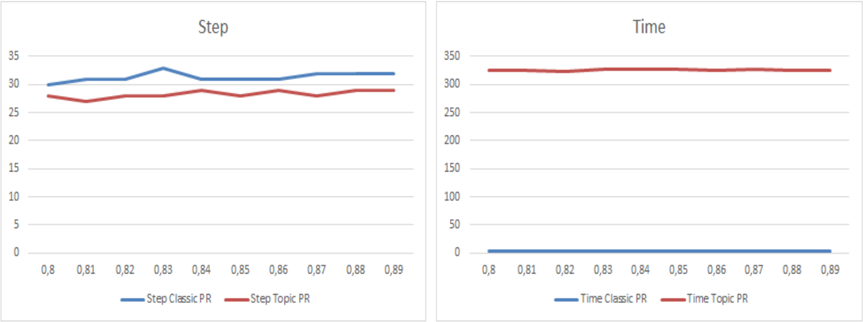
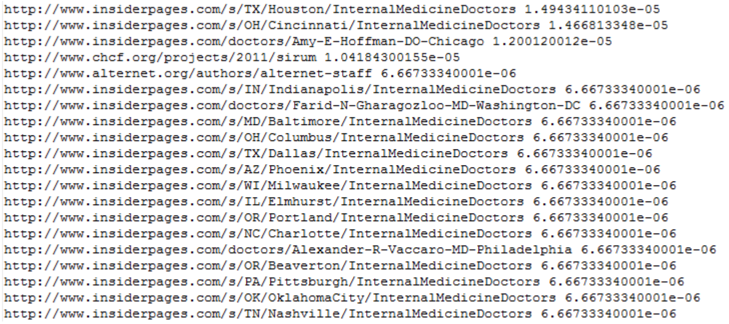


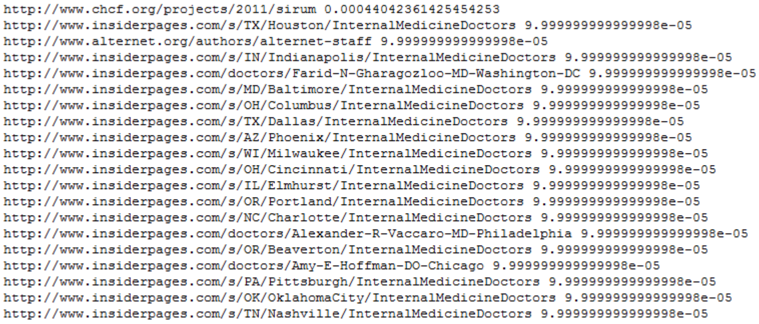
Figure 8: Times/steps of the combined algorithms

The last test was made to compare the output documents between the queries ”medicine” and ”medicine has existed for thousands of years”. First of all, we show the results of the first query. In particular, there are the outputs of basic best match and basic or topic page rank



```
http://www.insiderpages.com/s/TX/Houston/InternalMedicineDoctors 1.49434110103e-05
http://www.insiderpages.com/s/OH/Cincinnati/InternalMedicineDoctors 1.466813348e-05
http://www.insiderpages.com/doctors/Amy-E-Hoffman-DO-Chicago 1.200120012e-05
http://www.chcf.org/projects/2011/sirum 1.04184300155e-05
http://www.alternet.org/authors/alternet-staff 6.66733340001e-06
http://www.insiderpages.com/s/IN/Indianapolis/InternalMedicineDoctors 6.66733340001e-06
http://www.insiderpages.com/doctors/Farid-N-Gharagozloo-MD-Washington-DC 6.66733340001e-06
http://www.insiderpages.com/s/MD/Baltimore/InternalMedicineDoctors 6.66733340001e-06
http://www.insiderpages.com/s/OH/Columbus/InternalMedicineDoctors 6.66733340001e-06
http://www.insiderpages.com/s/TX/Dallas/InternalMedicineDoctors 6.66733340001e-06
http://www.insiderpages.com/s/AZ/Phoenix/InternalMedicineDoctors 6.66733340001e-06
http://www.insiderpages.com/s/WI/Milwaukee/InternalMedicineDoctors 6.66733340001e-06
http://www.insiderpages.com/s/IL/Elmhurst/InternalMedicineDoctors 6.66733340001e-06
http://www.insiderpages.com/s/OR/Portland/InternalMedicineDoctors 6.66733340001e-06
http://www.insiderpages.com/s/NC/Charlotte/InternalMedicineDoctors 6.66733340001e-06
http://www.insiderpages.com/doctors/Alexander-R-Vaccaro-MD-Philadelphia 6.66733340001e-06
http://www.insiderpages.com/s/OR/Beaverton/InternalMedicineDoctors 6.66733340001e-06
http://www.insiderpages.com/s/PA/Pittsburgh/InternalMedicineDoctors 6.66733340001e-06
http://www.insiderpages.com/s/OK/OklahomaCity/InternalMedicineDoctors 6.66733340001e-06
http://www.insiderpages.com/s/TN/Nashville/InternalMedicineDoctors 6.66733340001e-06
```

Figure 9: Output for basic best match / basic page rank

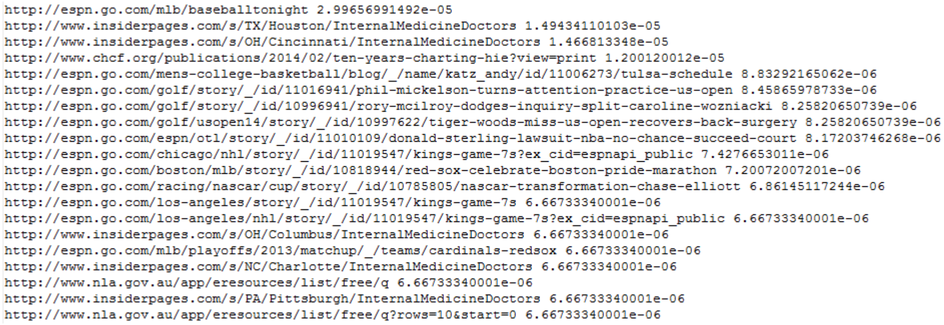


```
http://www.chcf.org/projects/2011/sirum 0.00044042361425454253
http://www.insiderpages.com/s/TX/Houston/InternalMedicineDoctors 9.999999999999998e-05
http://www.alternet.org/authors/alternet-staff 9.999999999999998e-05
http://www.insiderpages.com/s/IN/Indianapolis/InternalMedicineDoctors 9.999999999999998e-05
http://www.insiderpages.com/doctors/Farid-N-Gharagozloo-MD-Washington-DC 9.999999999999998e-05
http://www.insiderpages.com/s/MD/Baltimore/InternalMedicineDoctors 9.999999999999998e-05
http://www.insiderpages.com/s/OH/Columbus/InternalMedicineDoctors 9.999999999999998e-05
http://www.insiderpages.com/s/TX/Dallas/InternalMedicineDoctors 9.999999999999998e-05
http://www.insiderpages.com/s/AZ/Phoenix/InternalMedicineDoctors 9.999999999999998e-05
http://www.insiderpages.com/s/WI/Milwaukee/InternalMedicineDoctors 9.999999999999998e-05
http://www.insiderpages.com/s/OH/Cincinnati/InternalMedicineDoctors 9.999999999999998e-05
http://www.insiderpages.com/s/IL/Elmhurst/InternalMedicineDoctors 9.999999999999998e-05
http://www.insiderpages.com/s/OR/Portland/InternalMedicineDoctors 9.999999999999998e-05
http://www.insiderpages.com/s/NC/Charlotte/InternalMedicineDoctors 9.999999999999998e-05
http://www.insiderpages.com/doctors/Alexander-R-Vaccaro-MD-Philadelphia 9.999999999999998e-05
http://www.insiderpages.com/s/OR/Beaverton/InternalMedicineDoctors 9.999999999999998e-05
http://www.insiderpages.com/doctors/Amy-E-Hoffman-DO-Chicago 9.999999999999998e-05
http://www.insiderpages.com/s/PA/Pittsburgh/InternalMedicineDoctors 9.999999999999998e-05
http://www.insiderpages.com/s/OK/OklahomaCity/InternalMedicineDoctors 9.999999999999998e-05
http://www.insiderpages.com/s/TN/Nashville/InternalMedicineDoctors 9.999999999999998e-05
```

Figure 10: Output for basic best match / topic sensitive page rank

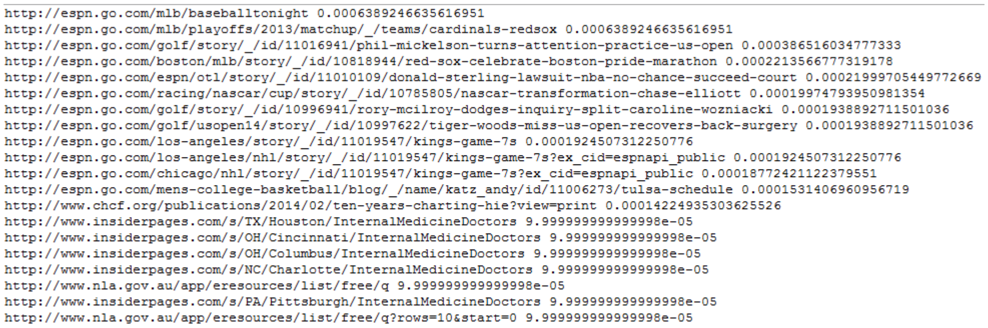
Now we show the results of ”medicine has existed for thousands of years”.

Afterward we analyzed which were the common documents of the two queries. Results show that there were five documents, and these are purely medicine pages. Furthermore, these five urls were the same for both basic and topic page rank.



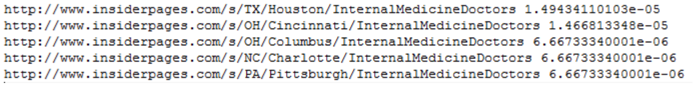
```
http://espn.go.com/mlb/baseballtonight 2.99656991492e-05
http://www.insiderpages.com/s/TX/Houston/InternalMedicineDoctors 1.49434110103e-05
http://www.insiderpages.com/s/OH/Cincinnati/InternalMedicineDoctors 1.466813348e-05
http://www.chcf.org/publications/2014/02/ten-years-charting-hie?view=print 1.200120012e-05
http://espn.go.com/mens-college-basketball/blog/_/name/katz_andy/id/11006273/tulsa-schedule 8.83292165062e-06
http://espn.go.com/golf/story/_/id/11016941/phil-mickelson-turns-attention-practice-us-open 8.45865978733e-06
http://espn.go.com/golf/story/_/id/10996941/rory-mcilroy-dodges-inquiry-split-caroline-wozniacki 8.25820650739e-06
http://espn.go.com/golf/usopen14/story/_/id/10997622/tiger-woods-miss-us-open-recovers-back-surgery 8.25820650739e-06
http://espn.go.com/espn/otl/story/_/id/11010109/donald-sterling-lawsuit-nba-no-chance-succeed-court 8.17203746268e-06
http://espn.go.com/chicago/nhl/story/_/id/11019547/kings-game-7s?ex_cid=espnapi_public 7.4276653011e-06
http://espn.go.com/boston/mlb/story/_/id/10818944/red-sox-celebrate-boston-pride-marathon 7.20072007201e-06
http://espn.go.com/racing/nascar/cup/story/_/id/10785805/nascar-transformation-chase-elliott 6.86145117244e-06
http://espn.go.com/los-angeles/story/_/id/11019547/kings-game-7s 6.66733340001e-06
http://espn.go.com/los-angeles/nhl/story/_/id/11019547/kings-game-7s?ex_cid=espnapi_public 6.66733340001e-06
http://www.insiderpages.com/s/OH/Columbus/InternalMedicineDoctors 6.66733340001e-06
http://espn.go.com/mlb/playoffs/2013/matchup/_/teams/cardinals-redsox 6.66733340001e-06
http://www.insiderpages.com/s/NC/Charlotte/InternalMedicineDoctors 6.66733340001e-06
http://www.nla.gov.au/app/eresources/list/free/q 6.66733340001e-06
http://www.insiderpages.com/s/PA/Pittsburgh/InternalMedicineDoctors 6.66733340001e-06
http://www.nla.gov.au/app/eresources/list/free/q?rows=10&start=0 6.66733340001e-06
```

Figure 11: Output for basic best match / basic page rank



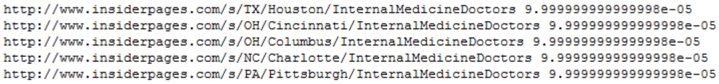
```
http://espn.go.com/mlb/baseballtonight 0.0006389246635616951
http://espn.go.com/mlb/playoffs/2013/matchup/_/teams/cardinals-redsox 0.0006389246635616951
http://espn.go.com/golf/story/_/id/11016941/phil-mickelson-turns-attention-practice-us-open 0.000386516034777333
http://espn.go.com/boston/mlb/story/_/id/10818944/red-sox-celebrate-boston-pride-marathon 0.0002213566777319178
http://espn.go.com/espn/otl/story/_/id/11010109/donald-sterling-lawsuit-nba-no-chance-succeed-court 0.00021999705449772669
http://espn.go.com/racing/nascar/cup/story/_/id/10785805/nascar-transformation-chase-elliott 0.00019974793950981354
http://espn.go.com/golf/story/_/id/10996941/rory-mcilroy-dodges-inquiry-split-caroline-wozniacki 0.0001938892711501036
http://espn.go.com/golf/usopen14/story/_/id/10997622/tiger-woods-miss-us-open-recovers-back-surgery 0.0001938892711501036
http://espn.go.com/los-angeles/story/_/id/11019547/kings-game-7s 0.0001924507312250776
http://espn.go.com/los-angeles/nhl/story/_/id/11019547/kings-game-7s?ex_cid=espnapi_public 0.0001924507312250776
http://espn.go.com/chicago/nhl/story/_/id/11019547/kings-game-7s?ex_cid=espnapi_public 0.00018772421122379551
http://espn.go.com/mens-college-basketball/blog/_/name/katz_andy/id/11006273/tulsa-schedule 0.0001531406960956719
http://www.chcf.org/publications/2014/02/ten-years-charting-hie?view=print 0.00014224935303625526
http://www.insiderpages.com/s/TX/Houston/InternalMedicineDoctors 9.999999999999998e-05
http://www.insiderpages.com/s/OH/Cincinnati/InternalMedicineDoctors 9.999999999999998e-05
http://www.insiderpages.com/s/OH/Columbus/InternalMedicineDoctors 9.999999999999998e-05
http://www.insiderpages.com/s/NC/Charlotte/InternalMedicineDoctors 9.999999999999998e-05
http://www.nla.gov.au/app/eresources/list/free/q 9.999999999999998e-05
http://www.insiderpages.com/s/PA/Pittsburgh/InternalMedicineDoctors 9.999999999999998e-05
http://www.nla.gov.au/app/eresources/list/free/q?rows=10&start=0 9.999999999999998e-05
```

Figure 12: Output for basic best match / topic sensitive page rank



```
http://www.insiderpages.com/s/TX/Houston/InternalMedicineDoctors 1.49434110103e-05
http://www.insiderpages.com/s/OH/Cincinnati/InternalMedicineDoctors 1.466813348e-05
http://www.insiderpages.com/s/OH/Columbus/InternalMedicineDoctors 6.66733340001e-06
http://www.insiderpages.com/s/NC/Charlotte/InternalMedicineDoctors 6.66733340001e-06
http://www.insiderpages.com/s/PA/Pittsburgh/InternalMedicineDoctors 6.66733340001e-06
```

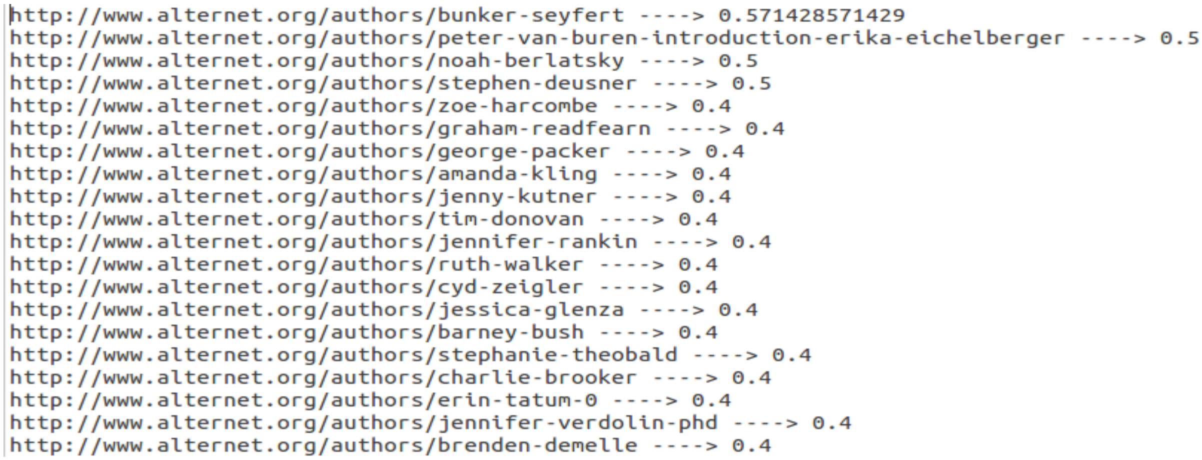
Figure 13: Common docs of basic page rank



```
http://www.insiderpages.com/s/TX/Houston/InternalMedicineDoctors 9.999999999999998e-05
http://www.insiderpages.com/s/OH/Cincinnati/InternalMedicineDoctors 9.999999999999998e-05
http://www.insiderpages.com/s/OH/Columbus/InternalMedicineDoctors 9.999999999999998e-05
http://www.insiderpages.com/s/NC/Charlotte/InternalMedicineDoctors 9.999999999999998e-05
http://www.insiderpages.com/s/PA/Pittsburgh/InternalMedicineDoctors 9.999999999999998e-05
```

Figure 14: Common docs of topic sensitive page rank

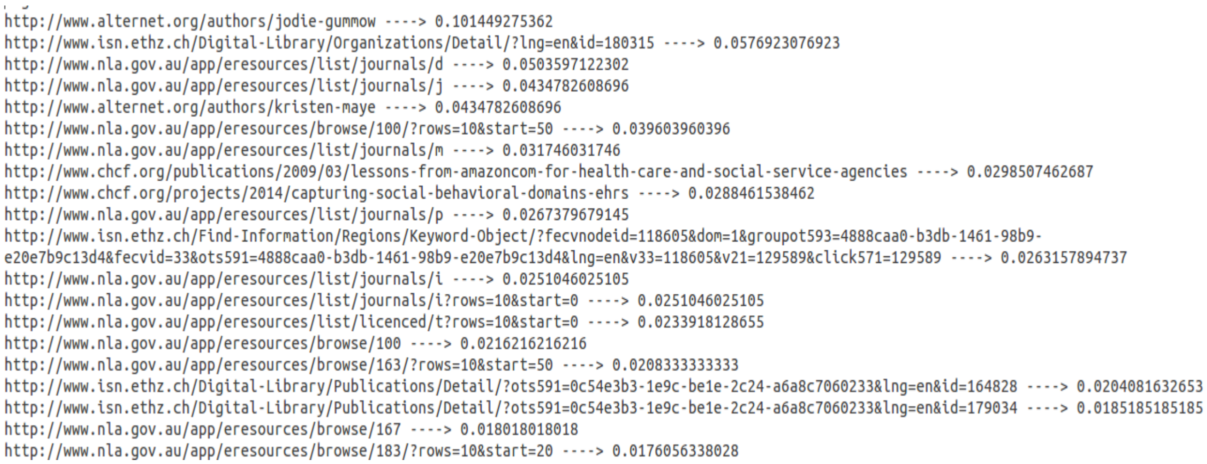
Finally, we runned a test on a very complex query, made up on several terms that we expected belong to different topics. This time we will show the result of both best match in order to show that the output of the optimized is completely different (it depends on the fist most popular words). The query is the following: *"The basic aim of the journal is to carry out researches on different disciplines of social sciences, more specifically on cultural and sport studies. Through the production of scientific researches, IntJSCS hopes to contribute to the internationalization process of ISCSA. Therefore the journal accepts researches in English language from all over the world."*



```
http://www.alternet.org/authors/bunker-seyfert ----> 0.571428571429
http://www.alternet.org/authors/peter-van-buren-introduction-erika-eichelberger ----> 0.5
http://www.alternet.org/authors/noah-berlatsky ----> 0.5
http://www.alternet.org/authors/stephen-deusner ----> 0.5
http://www.alternet.org/authors/zoe-harcombe ----> 0.4
http://www.alternet.org/authors/graham-readfearn ----> 0.4
http://www.alternet.org/authors/george-packer ----> 0.4
http://www.alternet.org/authors/amanda-kling ----> 0.4
http://www.alternet.org/authors/jenny-kutner ----> 0.4
http://www.alternet.org/authors/tim-donovan ----> 0.4
http://www.alternet.org/authors/jennifer-rankin ----> 0.4
http://www.alternet.org/authors/ruth-walker ----> 0.4
http://www.alternet.org/authors/cyd-zeigler ----> 0.4
http://www.alternet.org/authors/jessica-glenza ----> 0.4
http://www.alternet.org/authors/barney-bush ----> 0.4
http://www.alternet.org/authors/stephanie-theobald ----> 0.4
http://www.alternet.org/authors/charlie-brooker ----> 0.4
http://www.alternet.org/authors/erin-tatum-0 ----> 0.4
http://www.alternet.org/authors/jennifer-verdolin-phd ----> 0.4
http://www.alternet.org/authors/brenden-demelle ----> 0.4
```

Figure 15: output of basic best match algorithm on complex query

The result consisting of all alternet pages confirm that the basic algorithm retrieves general famous documents based on overall scores.



```
http://www.alternet.org/authors/jodie-gummow ----> 0.101449275362
http://www.isn.ethz.ch/Digital-Library/Organizations/Detail/?lng=en&id=180315 ----> 0.0576923076923
http://www.nla.gov.au/app/eresources/list/journals/d ----> 0.0503597122302
http://www.nla.gov.au/app/eresources/list/journals/j ----> 0.0434782608696
http://www.alternet.org/authors/kristen-maye ----> 0.0434782608696
http://www.nla.gov.au/app/eresources/browse/100/?rows=10&start=50 ----> 0.039603960396
http://www.nla.gov.au/app/eresources/list/journals/m ----> 0.031746031746
http://www.chcf.org/publications/2009/03/lessons-from-amazoncom-for-health-care-and-social-service-agencies ----> 0.0298507462687
http://www.chcf.org/projects/2014/capturing-social-behavioral-domains-ehrs ----> 0.0288461538462
http://www.nla.gov.au/app/eresources/list/journals/p ----> 0.0267379679145
http://www.isn.ethz.ch/Find-Information/Regions/Keyword-Object/?fecvnodeid=118605&dom=1&groupot593=4888caa0-b3db-1461-98b9-
e20e7b9c13d4&fecvid=33&ots591=4888caa0-b3db-1461-98b9-e20e7b9c13d4&lng=en&v33=118605&v21=129589&click571=129589 ----> 0.0263157894737
http://www.nla.gov.au/app/eresources/list/journals/i ----> 0.0251046025105
http://www.nla.gov.au/app/eresources/list/journals/i?rows=10&start=0 ----> 0.0251046025105
http://www.nla.gov.au/app/eresources/list/licenced/t?rows=10&start=0 ----> 0.0233918128655
http://www.nla.gov.au/app/eresources/browse/100 ----> 0.0216216216216
http://www.nla.gov.au/app/eresources/browse/163/?rows=10&start=50 ----> 0.0208333333333
http://www.isn.ethz.ch/Digital-Library/Publications/Detail/?ots591=0c54e3b3-1e9c-be1e-2c24-a6a8c7060233&lng=en&id=164828 ----> 0.0204081632653
http://www.isn.ethz.ch/Digital-Library/Publications/Detail/?ots591=0c54e3b3-1e9c-be1e-2c24-a6a8c7060233&lng=en&id=179034 ----> 0.0185185185185
http://www.nla.gov.au/app/eresources/browse/167 ----> 0.018018018018
http://www.nla.gov.au/app/eresources/browse/183/?rows=10&start=20 ----> 0.0176056338028
```

Figure 16: output of optimized best match algorithm on complex query

As we expected, the difference between the two algorithms is evident. The main point is that optimized best match gave in output documents of many different topics. The reason is that, if we consider the first words most recurrent across the documents, they may be very generic and with high probability they belongs to few other categories. For example we have the **isn** domain which refers to a youth academy probably belong to *kids and teens* topic. Moreover, the nla (National Library Australia) could refers to the *arts* topic. Finally, chcf.org certainly belong to the *health* topic.

Now let us consider the performances:

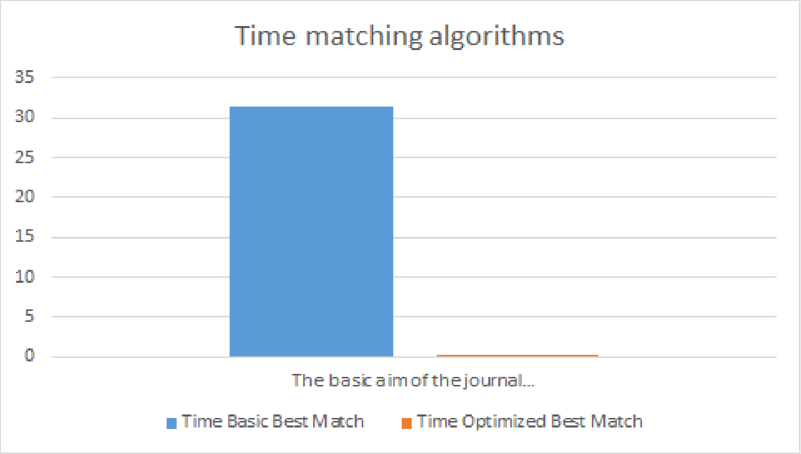


Figure 17: times elapsed for matching algorithms

It is clearly evident that the optimized algorithm is hugely more efficient than the basic one (little more one second for the first against the almost 33 seconds for the second). That is exactly what we expected if we consider that the optimized evaluates the scores only for a little threshold of documents.

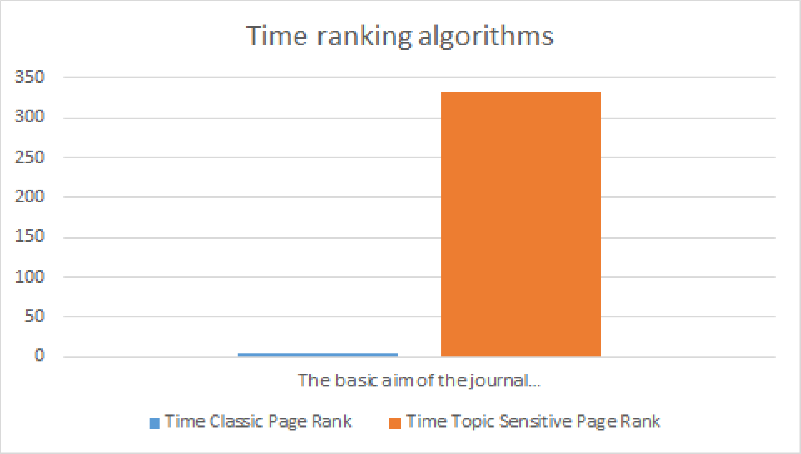


Figure 18: times elapsed for ranking algorithms

This graphics points out both ranking algorithms (e.g. for ß=0.8). It is clear that topic sensitive requires more time to converge than the basic algorithm.

That is predictable if we consider that in topic sensitive we have to compute as many rank vectors as the number of topics considered.

3.3 Conclusion

We have showed different results based on several query input inherent to both accuracy and time performance. We think that the two approaches should be used in different situations. Starting from matching algorithms, the basic best match is more efficient if we are interested in more accurate results. That is because the basic algorithm finds all the documents that, given a query term, have the highest overall score for the keywords it deals with. On the other hand, the times elapsed are not particularly performing. Instead, in the optimized algorithm we have a strongly reduced computing times. That is because we consider only K documents in which the first frequent query terms. However, this approach has a negative aspect too if we think about the accuracy. In fact we consider very generic words that appear in many different documents (which could belong to different topics). About the page rank algorithms, arise that the basic page rank is more performing than the topic sensitive if we only consider the time elapsed. That is because, topic sensitive has to compute different rank vectors for each topic, however, it requires less steps to converge to a solution.

4 Sponsored search

In this section we are going to talk about the sponsored search network, where advertiser text ads are shown on the result pages of user search queries. For each possible query, advertisers compete in auctions to determine the order in which the ads will be presented. According to this, the goal of auction mechanism is to assign a position (slot) to all of the advertisers competing for a specific query, based on their bid matching the amount they want to pay for each click. We will discuss about two auction formats:

1. First Price Auction
2. Vickrey–Clarke–Groves Auction

Those are used to decide the slots' winners and the amount of money they have to pay. Moreover, for each format, we will simulate some auctions using several bots, which are programs that, considering an history of previous auctions (to make it simple we only look at the last one), in order to suggest advertisers the best bid for the next auction following different strategies:

1. **Balanced best-response**: Evaluate which bid to choose for winning the desired slot following balanced tie-breaking rules;

   (a) *tie-breaking rule 1*: Returns a bid matching to the minimum value between the average of the sum of the evaluations of all slots and the last bid of the previous step (if he has enough budget, otherwise his bid is the remaining budget).

   (b) *tie-breaking rule 2*: Returns a bid matching the half of the sum between the advertiser's evaluation of the best slot and the previous winning bid for that slot, if he has enough budget, otherwise his bid is the remaining budget).

   (c) *tie-breaking rule 3*: Returns a bid matching approximately the value between the desired slots (suppose slot j) and the previous one (j-1), (if he has enough budget, otherwise his bid is the remaining budget). For example, I will bid a value closer to j-1, so I am indifferent from taking j at computed price or j-1 at price matching my bid.

2. **Competitor-busting best response**: Advertisers always submits the highest value for winning the desired slot:

(a) *tie-breaking rule 1*: Returns a bid matching the maximum between the average of the sum of the evaluations of all slots and the last bid of the previous step (if he has enough budget, otherwise his bid is the remaining budget).

(b) *tie-breaking rule 2*: Returns his value for the best slot if he has enough budget, otherwise his bid is the remaining budget).

(c) *tie-breaking rule 3*: Returns a bid matching of the previous bid for slot j-1 (if the desired slots is j) decreased of an epsilon value (if he has enough budget, otherwise his bid is the remaining budget).

3. **Altruistic best-response**: Advertisers always submits the lowest bid for winning the desired slot:

   (a) *tie-breaking rule 1*: It is the same of balanced bot.

   (b) *tie-breaking rule 2*: Returns a bid matching the minimum between the value for the best slot and the last winning bid for that slot incremented by an epsilon value (if he has enough budget, otherwise his bid is the remaining budget).

   (c) *tie-breaking rule 3*: Similarly to the second rule, returns a bid matching the minimum between the value for the desired slot (suppose j) and the last winning bid for that slot incremented by an epsilon value (if he has enough budget, otherwise his bid is the remaining budget).

4. **Competitor-bursting bot**: Always submits a bid greater than the highest bid seen in previous auction. In particular it returns the bid matching the value of the winning bid in the last step incremented by epsilon (if he has enough budget, otherwise his bid is the remaining budget).

5. **Budget-saving bot**: Always submits a value that is the minimum among the last non-winning bid and the advertiser minimum value for each slots.

6. **Random bot**: Always submits a random bid.

7. **Combination bot**: This bot is a combination of the above bots. For each query evaluates the strategy to apply considering three factors:

   - *current budget*: The remaining budget of the advertisers
   - *state of auction*: How many query's slots have already been assigned

- *advertiser value*: The average evaluation of the advertiser for all the query's slots

Keeping in mind these factors we distinguish four main cases:

(a) *Case 1*: Low value for the slot. For instance, I have a small consideration of that item and I don't want bid too much even if the auction is almost finished. For this reason with high probability I will play altruistic or budget saving.

(b) *Case 2*: Fairly good value but the auction has just begun. For instance, I have a quite good consideration of that item but I wouldn't submit a high bid if the auction is on first steps. We have to make distinction of two sub-cases:

   i. *Sub-case 1*: I would bid higher if my current budget is high enough (i.e. higher than half of the starting budget). In this case I will play a competitor bursting strategy more likely.

   ii. *Sub-case 2*: I wouldn't bid higher because my current budget is not high enough. So more likely I will play a budget saving strategy.

(c) *Case 3*: Fairly good value and the auction is almost finished. For instance, I have a quite good consideration of that item and I would submit a high bid if the auction is on last steps. Also in this case we have to consider the same two sub-cases seen before (based on the current budget of the advertisers), with the difference that even more probably the players will play a competitor bursting approach (go for broke strategy).

(d) *Case 4*: High value for the slot. In this case if I strongly want that item I will chose a high bid without considering remaining budget or the state of the auction. So with probably nearby 1 I will play a competitor bursting strategy.

As well as the cases just shown, we want also take account of people who may not follow reasoning. We call it *"crazy bidder"* and they would play with very low probability a random strategy, no matter the factors seen above.

Afterwards, we will show the performance of each auction format considering the following factors:

1. **Advertisers utility**: which is the difference between the value of the advertiser for the slot won and the product of his payment and the click-through rate value.

2. **Company revenue**: which is the sum, for each slot, of the winner advertiser's payment and the click-through rate value of the slot.

Finally. we perform a comparison, upon this values, between the bots in order to understand which one guarantee the highest revenue values.

5 Testing approach

Firstly, we have done a simple static test in order to understand the overall behaviours of every both we have implemented. Afterwards, we tried different dynamic tests whose input is randomly generated for each run.

5.1 Static tests

We runned the static test for both the auction formats with the following setting parameter:

- 2 queries
- 3 advertisers
- 3 slots for each query

The slots values of the advertisers and the click-through rates has been chosen fixed in a range (1-10), whereas the starting budget of the advertisers was a fixed value between the range (60 -150). The figure 3 shows the behaviour of the bots if every advertiser use the same strategy:

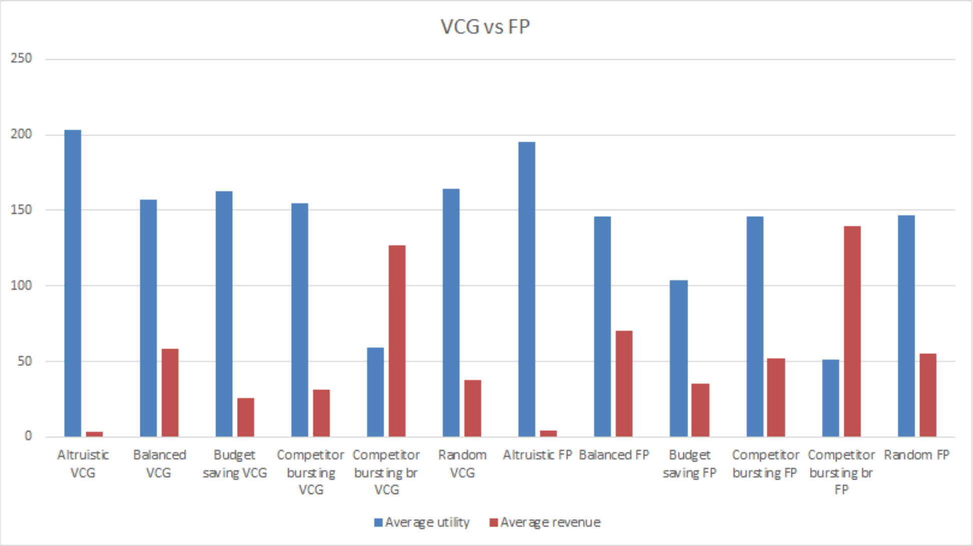


Figure 19: Utilities and revenues of first static test

We can clearly see that the average utility of the advertisers is higher if it is calculated according to VCG, that was exactly what we expected. On the other hand, the overall revenue of the company is surely better if the payments are computed with the First Price approach. In both the auction formats is evident that the balanced and the competitor bursting guarantees the highest revenue for the company, otherwise the highest utilities (especially for VCG) are provided from the altruistic, balanced and budget saving.

The following graphic points out a different situation: now we have a single advertiser which has a fixed behaviour (he always plays the same strategy), whereas the other players have different attitudes (they follow the combination heuristic). We planned this test to understand the tendency of the advertisers' behaviour in a more realistic situation:

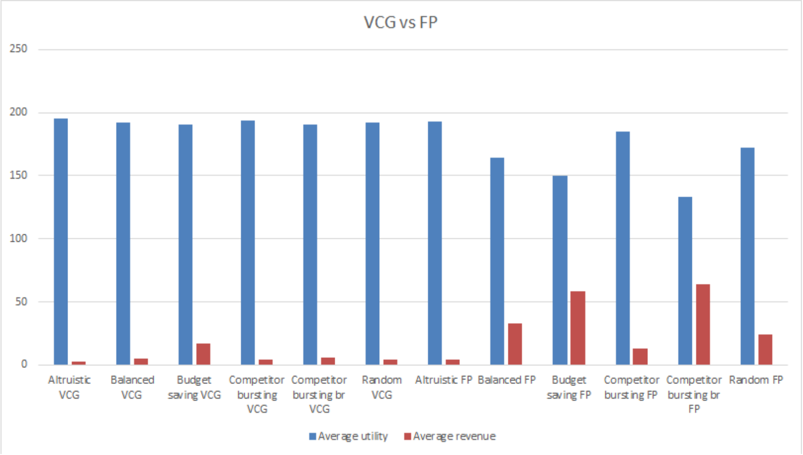


Figure 20: Average utility and revenue of second static test

The overall output is almost the same of the previous case, although the values are more balanced because the advertisers play opposite strategies. We can also notice an anomaly occurred during the budget saving run because we expected high utility as well as low revenue. The reason of this anomaly is that: even if the static advertiser submits the lowest bid, the others mostly bid the highest (according to the competitor burst). This leads to high payments and, consequently, high revenue for the company.

To enforce what we have said so far, we now point out the graphics of the averages value for revenue and utility in both type of auctions:

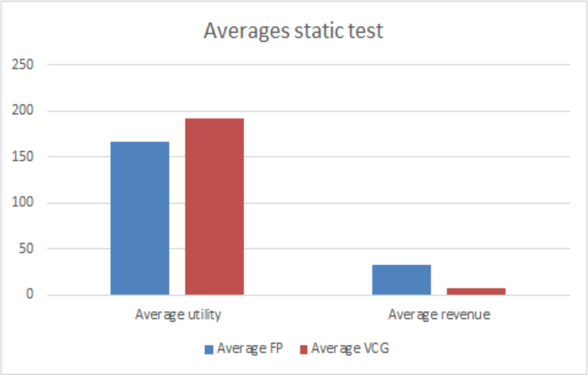


Figure 21: Overall average of performances for FP and VCG auction

5.2 Dynamic tests

Once we have seen how the players deal each other, we want show the performance of every bot in a ”dynamic” context. First of all, we compare the bots each other in order to realize which one gives the best values of utility and revenue. Each test is based on 500 iterations in order to obtain the average values.

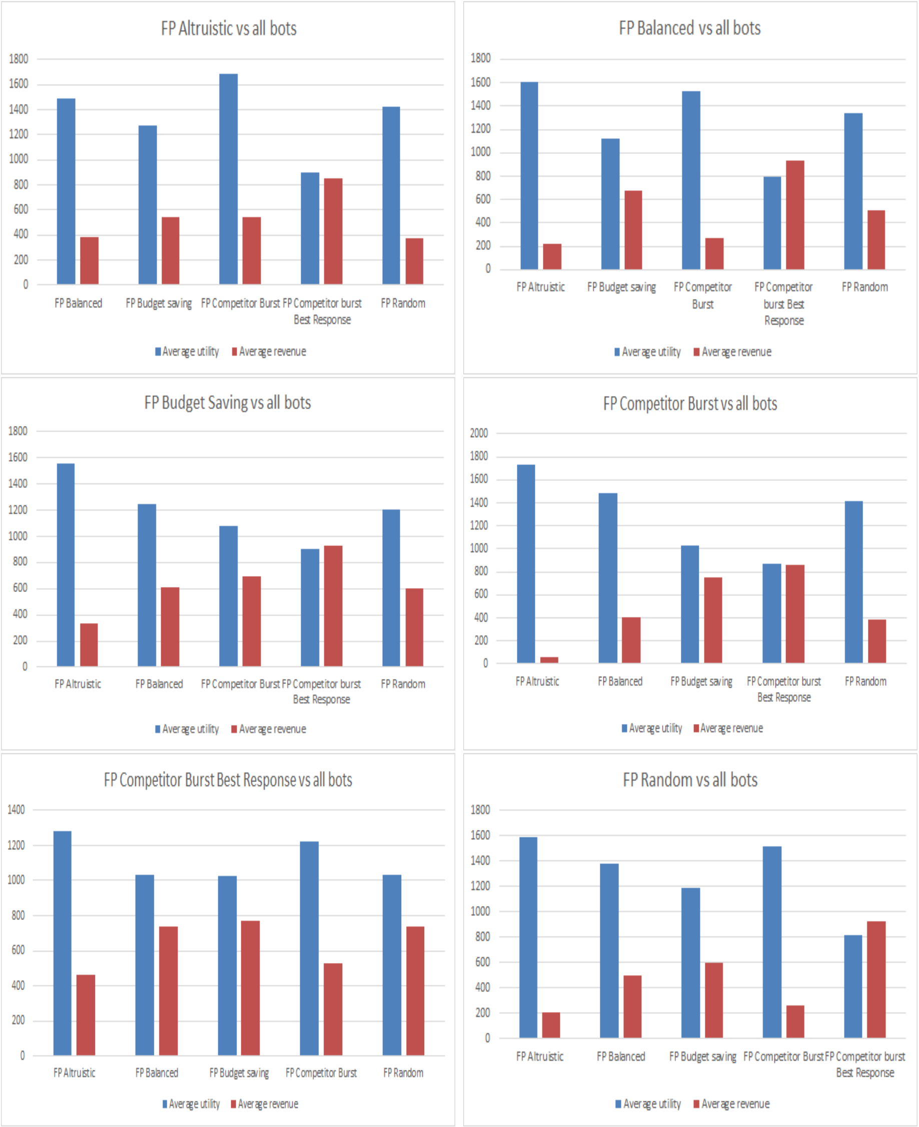


Figure 22: Comparison between the FP bots

At first glance, except for the random bot, as we could imagine, it seems that the competitor burst best response bot provides the highest revenue values. On the other hand we have unexpected high revenue results using the budget saving strategy.

Analyzing the values obtained by the competitor burst best response vs the others, it is evident that the highest revenue value occurs when it plays against the budget saving (almost 800). It could be acceptable if we think that the latter bid the minimum value between his medium evaluation and the lowest bid previously seen. It happens that the minimum value of a high bid previously submitted and a great evaluation is still a high bid which budget saving players submit. The second highest revenue instead is given by the interaction with the balanced bot. We initially expected major values from the interaction with the competitor bursting best response, but it gave only a value between 400 and 600. We think that the best response always plays the third tie breaking rule. For this reason the bids are not as high as we expected and, moreover, the best response bot raises them only of a small amount (epsilon=0.1). On the other hand, talking about the utility, it is clear that an altruistic approach grants the better values, even though the balance bot performs good in this sense as well.

Now we are going to display the results given by the bots with the VCG auction format:

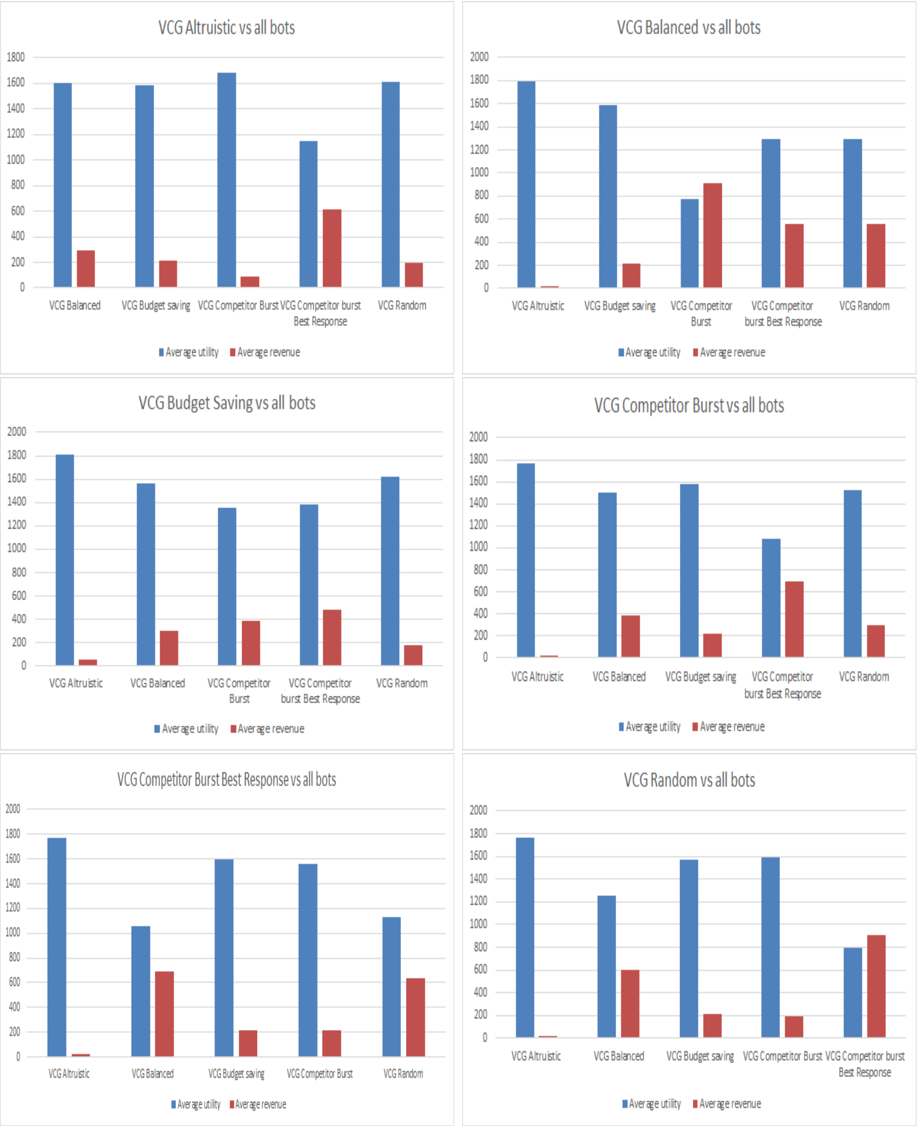


Figure 23: Comparison between the VCG bots

At first glance, as we hope, the VCG format strongly promotes the utility of the advertisers (it seems that altruistic budget saving bots provide better performance). Looking at the revenue it is obvious that the values are fairly low. It is evident that the highest rates occur when both competitor burst type and the balanced bots deal each other. So, in conclusion we can consider them as the best for the agency.

As long as, we have seen the general behaviour of the bots for the two different auction formats in various situations, we will now show the overall performance in a very realistic auction game. We ran all the bots according to the heuristic described above. The input has been chosen randomly for each of the 500 repetitions of the game. Furthermore, we repeated the experiment two times: the first time we chose a lower threshold, according to which player decides to adopt a competitor bursting strategy; the second time we augmented slightly that threshold in order to evaluate the impact on the general trend of the auctions (we expected that players would play a competitor bursting strategy more rarely, it follows that the overall revenue should be lower, whereas the utility should be higher).

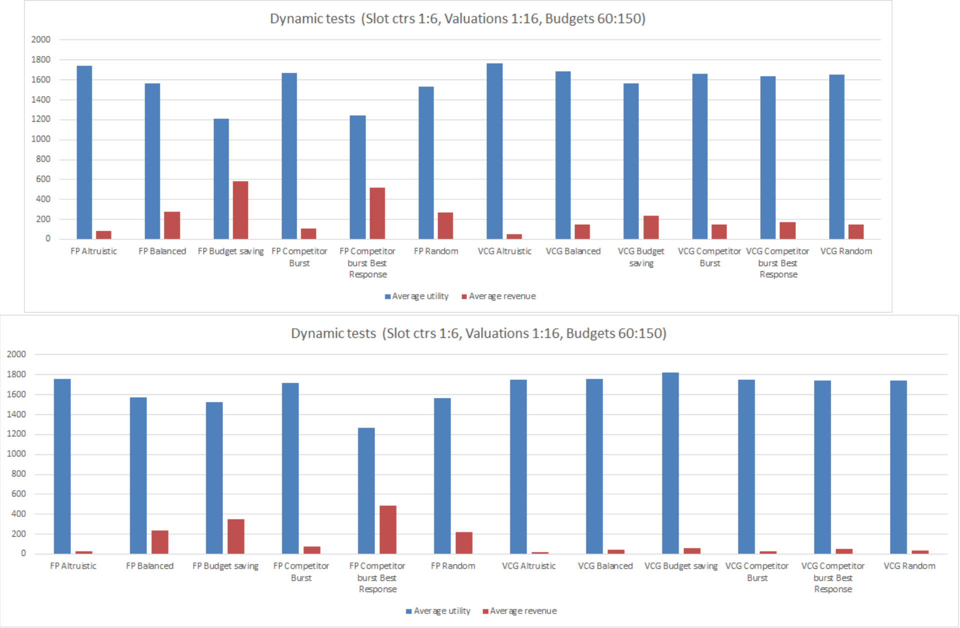


Figure 24: Average utility and revenue in a realistic auction game

As we supposed, decreasing the number of times an advertiser use a bursting strategy, the revenue decreases significantly beyond the kind of bot we are considering.

Considering that the bots which guarantee highest revenues are balanced, competitor bursting and budget saving, we are going to prove which one, for each auction format, most satisfies the advertisers in terms preferred slots assigned. Assuming that each advertiser applies the same strategy, the following are the average result values of 500 repetitions:

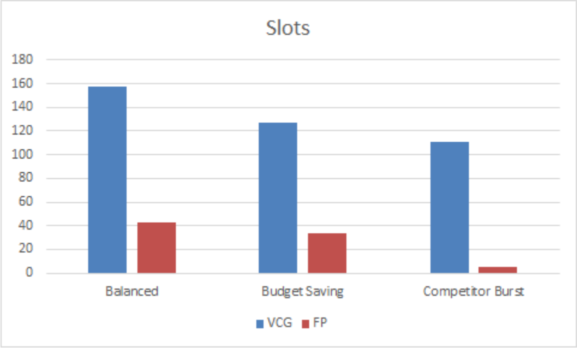


Figure 25: Preferred slots assigned for first price and VCG bots

We notice that the values are highly different between the two auction format. This is explained by the fact that the VCG have an average step number greater than the First Price format (we only print the average of the 500 repetitions of the tests).

It is evident that the balanced bot is the best for each auction format. We have also good performance from the budget saving. On the other hand the competitor bursting approach provides low performance in this sense.

We have also tried a test using the combination approach, the results are the following:

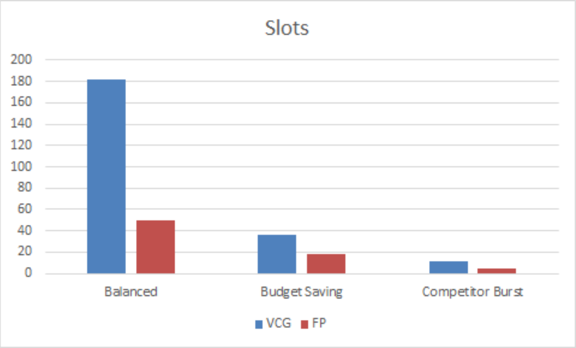


Figure 26: Preferred slots assigned for first price and VCG bots

The trend is the same of the last case, however it strongly depends on the fact that balanced is the mostly played strategy in a random auction game.

5.3 Conclusions

For all the reasons that we have already explained, we could assert that the best bot depends on what we want to consider:

- *From the advertisers point of view*, we should consider both the utility and the number of times they win the desired slots. For the utility we think that the best approach is the altruistic. On the other hand, we have showed that the balanced bot is the best strategy for winning the slot which mostly satisfies the players.
- *From the company point of view*, the best bot certainly is the competitor bursting best response, because independently from the other both whereby compete, it always provides the highest revenue.