2009 IEEE International Conference on Semantic Computing

If you ask nicely, I will answer: Semantic Search and Today's Search Engines

Tomasz Imielinski
Ask.com R&D
Edison, NJ
timielinski@ask.com

Alessio Signorini
Department of Computer Science
University of Iowa
Iowa City, IA
alessio-signorini@uiowa.edu

***Abstract*—Today's search engines are still very sensitive to the way queries are constructed. In some occasions, equivalent but slightly different forms of a query lead to completely different results. However, popular queries with only one right answer seem to be generally well served by search engines, which generally return the correct answer among their top 10 search results. Internet's redundancy of information and the recent proliferation of user generated content helps search engines to remain almost entirely keyword oriented and still robustly handle equivalent versions of queries. In this paper we propose a family of metrics to evaluate the semantical invariance of a given search engine, and we report experimental results for well-known engines such as Google, Yahoo!, Live and Ask.com, as well as for new semantic search engines like Hakia and Cuil.**

***Keywords*-Semantic Search; Rephrasing Invariance; Results Stability; Search Engines;**

I. Introduction

In the last years many "semantic search engines" have been launched and some often use this word while comparing themselves with others. Nowadays, the lack of proper metrics to assess "how semantic" a search engine allows for any claim to be made, and the word "semantic" gets often abused while looking for quick fame. But what makes a search engine "semantic"? Is the use of some Natural Language Processing (NLP) technologies sufficient to award such definition? When can a user say that the search box it has in front is really understands its queries in the same way another human would?

Imagine the following dialog between user and a search engine:

> User: How is the weather in Hawaii?
> *Engine*: I do not know.
> User: What is weather in Hawaii islands now?
> *Engine*: I already told you, I dont know it.
> User: current weather in Hawaii, USA.
> *Engine*: Hye, I told you already, I have no idea.

This search engine is clearly clueless about the status of the weather in Hawaii but it is nevertheless semantic: it knows that it does not know. Notice also how it understands that the user keeps asking the *same* query, although differently phrased.

It is important to keep in mind that even the best and most complete semantic search engine might not have all the answers: it is possible for such an engine to have all the knowledge of the world in its database and not be semantic at all, or have very little knowledge (or even none!) and be truly semantic.

Today's search engines often return different answers for variations of the same query. Those answers may be right or wrong depending on the keywords used in the query. The burden of selecting the right keywords is left to the user which often will have to formulate its query multiple times to obtain its answers. In this aspect, search engines unfortunately make very little effort to help users with the task, potentially missing good opportunities.

In this paper we propose several metrics to measure "how semantic" a given search engine is and we present numerous experimental results. In our experiments we compared well-known search engines such as Google, Yahoo!, Live and Ask.com, with Hakia and Cuil, two new search engines which claim to be semantic.

A. The Human Search Engine

The results returned by a truly semantic search engine should be invariant to the way the query is formulated (rephrase). As shown in the introductory dialog, there is no benefit in reformulating the question since the engine understands that the user is simply restating the same one, for which it has no answer.

Identifying semantic equivalence among queries is fairly easy for humans but unfortunately very hard for automated systems. The ideal semantic search engine would emulate a super-human, an hypothetical Human Search Engine (HSE) with a memory big enough to keep at hand all the questions previously asked and the answers given. For any new question equivalent to one previously heard, the HSE will quickly recall the earlier answer and provide it again as a response. Different HSEs could give different answers to the same question, depending on their knowledge and intelligence, but

978-0-7695-3800-6/09 $26.00 © 2009 IEEE
DOI 10.1109/ICSC.2009.31

each of them would preserve its invariance with respect to queries which are semantically equivalent (i.e. have the same meaning).

B. Semantic Invariance

Achieving *semantic invariance* is a necessary condition for any search engine which aims to be defined "semantic", and while this is implicit in the HSEs example mentioned above, today's web search engines are still far away from this goal.

In a semantic search engine, two queries which are semantically equivalent should return the same results presented in the same order. Users expect this kind of behavior from human operators and search engines should imitate that. For example, the queries "biography of George Bush", "bio of George Bush" or "find me bio of George Bush" should always return the same results when submitted to a semantic search engine.

However, when there is only One-Right-Answer (ORA) for a query, this condition can be less restrictive since we are only interested in the presence of the answer: if the SEarch Result Page (SERP) contains the answer for one query, it should also contain the answer for all its semantically equivalent rephrases. Under this weaker definition of equivalence, the SERP for "capital of France" and "which city is France's capital" should both contain the answer "Paris".

Every search engine which claims to be semantic has been (or will be) challenged by some users which will try to find equivalent rephrases (with no answer) for their questions. Semanticly equivalent queries should always return the same answer, if the search engines knows it, or none at all. For example, it would be an error to return the correct answer for both of rephrases above while asking for France, but only for the first one when the target country is England. Conversely, different queries (e.g., "where is Michelle Obama?" and "what is Michelle Obama?") should not return the same results just because they share the subject.

Invariance with respect to ORA queries is less restrictive than overall semantic invariance since it allows both different results and ordering, as long as the answer is in the result page. In the rest of this paper we adopted such weaker definition while presenting some of the experimental results, although we firmly believe that the SERP of a truly semantic search engine should not be affected by the particular rephrasing of the query.

C. Invariance Metrics

In this section we introduce some of the invariance metrics which will be referenced throughout the rest of the paper. The following measures try to capture the semantic sensitivity of each search engine.

Entropy As said in the introduction, an ideal semantic search engine should always return the same set of results, in the same order, for equivalent queries. The stability of the SERP is probably the most significative sign of true semantic understanding of the query. To measure the stability of the results we computed the Entropy, a well-known statistical measure [1] in Information Theory.

The following is a more formal definition of the concept, applied to SERP stability: let $Q = q_1, q_2, ..., q_n$ be a set of equivalent query rephrases, we define $p(u, Q, k)$ as the probability that URL u will be returned in position k as result for queries in Q. For example, given a set of equivalent queries Q and a URL u, then $p(u, Q, 1)$ is the probability that such URL will be returned in top position, $p(u, Q, 2)$ in second position, and so on. In a perfectly semantic search engine, $p(u_i, Q, k_i) == 1$ for all the values of i. If the set Q contains N equivalent queries, then the stability $p(u, Q, k)$ of each URL u will be a value from the set $1/N, 2/N, ..., N - 1/N, 1$, and since $\sum_{u \in URLs} p(u, Q, k) = 1$ it satisfies the requirement for probability distribution.

Thus, given a set of equivalent rephrases Q and a position k, it is possible to calculate the entropy of the dataset as $Entropy(Q, k) = -\sum_{u \in URLs} p(u, Q, k) * log_2 p(u, Q, k)$. Since more than 65% of the search clicks are done on the first result, during our experiment we considered only the top position ($k = 1$). The entropy of a each search engine has been computed as the average of the entropy calculated on each dataset.

Top-K Results Overlap Given two semantically equivalent queries r_1 and r_2, this measure aims to compute the fraction of URLs which are shared among their top K results. Ideally, a semantic search engine would return the same results for both instances making the overlap stable to 100%.

A more formal definition of the measure is the following: let q_1 and q_2 be two queries and let $u_1, ..., u_P$ be the URLs which are returned among top K results by the search engine for both queries, then $Overlap(q_1, q_2, K)$ is defined as P/K. During our experiments we will be measuring the pair-wise overlap for queries which are semantically equivalent

ORA Invariance Given a set Q of semantically equivalent queries $q_1, ..., q_n$, this test computes the fraction of queries of Q for which the correct answer appears in the result page. Assuming that the engine has the knowledge necessary to answer, due to the invariance with respect to the query form the result of this test should be stable to 100% in any truly semantic search engine.

D. Query Schemata

Manually generating the 40,000+ queries used during our experiments would have been, if at all feasible, very time consuming and highly subjective. For this reason, we decided to rely mostly on popular templates extracted from

AOL's query logs, parameterized with entities extracted from pre-classified Wikipedia pages.

Query Schemata are query templates which are parameterized by classes of objects, for example: "bio of *person*", "President of *country*" or "side effects of *drug*" are all valid query schemata. Real queries can be generated by substituting the free parameter with any element of the corresponding object class, such as "bio of George Bush", "President of Spain" or "side effects of Tylenol". Query schemata are not restricted to natural language questions and may include any type of query.

Query Schema A query schema is denoted by $Q(d_1, d_2, ..., d_N)$ where $d_1..d_n$ are entity belonging to class D. Two query schemata $Q_A(D)$ and $Q_B(D)$, where D is a class of entities, are semantically equivalent if for any member of the class D the resulting queries have exactly the same meaning and are expected to return the same answers/results when submitted to the same Human Search Engine (HSE). Generally speaking, any schema Q_x with the same property can be considered a "rephrase" of Q_A, and we will reference to $[Q]$ as the union of all the equivalent Q_x schemata.

For example, the two query schemata Q_Y ("What is the weather of *country*?") and Q_W ("How is *country*'s weather?") are semantically equivalent since they expect to return the same results for any element of the class *country* (as well as for cities, states, and so on).

In the rest of the paper we will be dealing with schemata with only one class only although the discussion carries in a straightforward way to schemata which support multiple classes, for example "When is the next concert of *singer* in *city*?" and "next *singer* concert in *city*" are equivalent schemata which accept two entities from differentbe classes.

II. Related Works

The academic community produced a considerable amount of research on Semantic Search over the past few years. At the time of writing, there are 379 papers published about "semantic search" according to Google Scholar[1] and 110 according to CiteSeer[2].

While many initiatives aim to make the Web semantic, the most authoritative is probably the W3C Semantic Web Activity [2] which aims to provide a common framework (based on the Resource Description Framework [3]) and allow data to be shared and reused across application, enterprise, and community boundaries. It is a collaborative effort led by W3C with participation from a large number of researchers and industrial partners. In their 2001 article [4] on the Scientific American Magazine, Berners & al. described for the first time a futuristic world in which the semantic web would allow deep interactions between different structured data.

Assuming that data will at some point be available in a well-defined structured format, many researchers focused their work on its retrieval. Guha, Mccool and Miller introduced in [5] an application called "Semantic Search", which aims to improve traditional web search exploiting correlated contents and their relationships. Cohen & al. presented in [6] XSEarch, a semantic search engine for XML. Their engine extends the way normal web search engines deal with XML separating unrelated pieces of XML before conducting the search to avoid false matches. In [7] Rocha suggested a search architecture that combines classical search techniques with spread activation techniques applied to a semantic model of a given domain. Also Bergamaschi and Guerra, in their SEWASIE project [8], aimed to design and implement an advanced search engine which would enable intelligent access to heterogeneous and distributed data sources through community-specific multilingual ontologies.

Being able to correctly understand and associate together different formulations of the same query/question is one of the hardest challenges in semantic search. Researchers participating to the TREC-QA track developed in the last years many methods for automatically creating rephrases of the same question. Brill & al. in [9] used simple words permutation to produce massive amounts of (mostly ungrammatical) rephrases for their questions. In [10] Lawrence and Giles used manual reformulation rules to generate rephrasing of the questions. Similarly, Kwok & al. used [11] transformational grammar to perform syntactic modifications of the questions.

Since manual reformulation and transformational grammar require intensive manual effort, some authors tried to exploit web search engines to automatically rephrase questions and queries. In [12] Duclaye & al. used the web to obtain different verbalizations for a seed relation (e.g., Author/Book). More recently, in her master thesis [13] Anna Hedstrom analyzed query logs to learn to generate new rephrases for its queries.

At the best of our knowledge, however, nobody tried to formally define metrics to asses how "semantic" a search engine is.

III. Experimental Results

In the following sections we will summarize the methodology adopted and the results obtained for some of the experiments conducted during this study.

A. Do you know that Picasso is a painter?

The catchy title of this section points out a common problem of today's search engines: over-specifying the query might actually hurt the relevance of the results returned. While

[1] http://scholar.google.com
[2] http://citeseer.ist.psu.edu

some additional description of the subject of the conversation usually helps in human dialogs, keyword based search engines will not identify its redundancy and will try to use it during text matching.

Such additional description could be for example the category of the object, and adding it to the query should not change the SERP returned. Submitting "IBM the company" instead of "IBM", "France the country" instead of "France" or "Tom Hanks the actor" instead of just "Tom Hanks", should not change the set of results returned by the search engine. This should also hold true for those queries that have multiple meanings (e.g. "Paris"), but for which the dominant one is so strong and well-known to eclipse the others.

To obtain the initial list of objects to over-specify, we extracted from Wikipedia all the pages which do not contain any disambiguation link and take advantage of an Infobox. The name of the Infobox has been used during the experiment as the additional keyword in the query. We decided to focus our tests around 9 main categories: actor, artist, city, company, disease, magazine, MLB player, politician and president.

Following this methodology we obtained a list[3] of more than 36,000 total entities to be used in our tests, which we reduced to about 8,000 imposing a limit of maximum 1,000 items for each category. Examples of entities included in our experiments are:

- Artists: Lyubov Popova, Antonio Zanchi, ...
- Companies: Iomega, Duracell, NetZero, Asus, ...
- Cities: Lakhva, Adahuesca, Fabletown, ...
- Diseases: Osteoarthritis, Diastrophic dysplasia, ...

We submitted the queries to both well-known engines (i.e. Google, Ask.com, Yahoo! and Live) and "semantic" search engines (i.e. Hakia and Cuil). Every query has been submitted both in its normal version and in an augmented one which included the Infobox category. For example, we had "Ashton Kutcher" and also its augmented version "Ashton Kutcher actor". Figure III-A shows the fraction of query pairs for which the top K overlap is 100%.
This graph summarizes how the results returned change, on each search engines, when the query is over-specified with the use of category information. For example, the top 1 result in Yahoo! is the same only for 45% of the query pairs, about 40% in Live results, 27% in Google's and about 22% in Ask.com search results. These numbers indicate that search engines mostly do not understand that Iomega is a company or that Richard Dieberkorn is an artist.

The addition of those superfluous terms resulted in different sets of pages emerging among the URLs returned. It is important to notice how this experiment does not aim to assess the difference in quality among those results, but rather tries to point out how results which should not differ in fact do. The fraction of query-pairs invariant to the addition of redundant category information is on average less than 5% considering top 3 results, and practically zero for all the search engines if more than 4 URLs are considered.

This shows an heavy reliance on tags and anchor text of pages by the search engines, which level of susceptibility varies between different pages and different objects.

[3]All the data and queries used in the experiments are freely available in a companion website

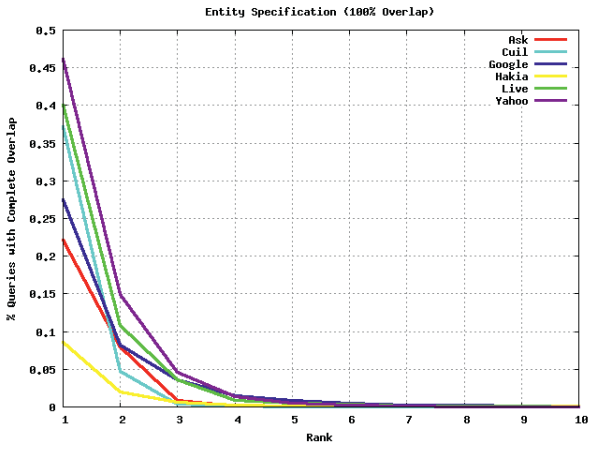

Figure 1. Fraction of query pairs for which the top K overlap is 100%

B. Are "top 10 songs" == "top ten songs"?

Our second experiment aimed to measure the sensitivity of search engines to the use of synonyms. However, instead of developing complex context-aware replacement algorithms which correctly understand when is appropriate to substitute a word with another, we decided to focus on a simple number transliteration. Working with numbers, instead of generic synonyms replacement, allowed to cheaply but accurately generate massive amounts of equivalent queries to be used during our tests. In addition, we have chosen this test because it is very simple, easy to understand, and simulates well a very common (and known) user behavior.

Our initial set of queries has been built extracting from AOL's query logs more than 1500 queries starting with "top *number*", and replacing the number with the word numeral (e.g. "top 20 cars" to "top twenty cars"). Examples of such queries are "top 1000 baby names", "top 100 games of all time", "top 20 rock hits" and "top 10 electric cars".

Table I and Table II summarize the fraction of query pairs which had respectively 100% and 0% URLs overlap among their top K results:

	Ask	Cuil	Google	Hakia	Live	Yahoo
1	0.029	0	0.057	0.125	0.025	0.025
2	0.004	0	0.018	0.040	0.007	0.004
3	0.004	0	0.004	0.018	0.004	0
4	0	0	0	0.015	0.004	0
5	0	0	0	0.007	0.004	0
6	0	0.004	0	0.007	0.004	0
7	0	0	0	0	0.004	0
8	0	0	0	0	0.004	0
9	0	0	0	0	0.004	0
10	0	0	0	0	0.004	0

Table I
FRACTION OF PAIRS WITH 100% OVERLAP AT DIFFERENT K

	Ask	Cuil	Google	Hakia	Live	Yahoo
1	0.971	1.000	0.943	0.875	0.975	0.975
2	0.957	1.000	0.883	0.740	0.949	0.950
3	0.946	1.000	0.799	0.623	0.935	0.922
4	0.925	1.000	0.689	0.542	0.913	0.900
5	0.903	1.000	0.633	0.509	0.909	0.883
6	0.878	1.000	0.604	0.484	0.891	0.872
7	0.867	1.000	0.601	0.476	0.891	0.851
8	0.853	0.996	0.590	0.462	0.884	0.833
9	0.842	0.993	0.587	0.451	0.865	0.819
10	0.832	0.986	0.576	0.436	0.855	0.819

Table II
FRACTION OF PAIRS WITH 0% OVERLAP AT DIFFERENT K

All the search engines examined did very poorly on this test. On average, only 3% of the pairs had any overlap at all even considering only the top URL returned. An exception is Hakia, which demonstrated a better (although still far from what was expected) understanding of this type of queries. An ideal semantic search engine should have had 100% overlap among all top 10 results for such trivial synonymous query pairs.

C. Do search engines understand rephrasing?

The goal of this set of experiments is to evaluate the invariance of the answer/results returned by different rephrasing of the same query. This experiment is divided in 2 parts: ORA Invariance and Stability.

To the first category belong query schemata which expect to find the answer in the result page. In those experiments we used many different rephrases: 28 for "Population of *country*", 11 for "ceo of *company*", 2 for "Author of *book*", 17 for "Super Bowl *year* Winner", 22 for "Wimbledon *year* Winner" and 17 for "World Cup *year* Winner". Those patterns have been harvested in AOL's query logs and manually parameterized. The correct answers have been manually found on publicly accessible web sites. While looking for accuracy the population of a country, some normalization have been applied to the numbers in the search page (e.g. "7.5M" was converted to "7500000") and the outcome was considered positive if there was at least a number in the page in the range of +/- 10% of the expected value.

In the second part of the experiment we were mostly interested in measuring the difference among the results returned for small modification of the original query. For example, we extracted 406 queries with the form "how to cook..." from AOL's query logs, manually checked them for consistency, and then parameterized the verb alternating between "cook", "make" and "prepare". Similarly, we obtained the list of the 100 Most Influential Books of all times from Wikipedia and manually parameterized the queries "who is the author of *book*" and "who wrote *book*" using such list. Finally, we obtained a list of common diseases from WebMD and collected more than 400 equivalent queries containing "symptom" or "sign" from AOL's query logs, which were parameterized using those diseases.

ORA Invariance In this test, for each query submitted to a search engine, we looked for the presence of the answer anywhere in the result page. If the answer was found, it was considered a positive outcome, otherwise, a negative one. In the end, for each of the ORA schemata we measured the fraction of rephrases for which each given search engine provided the correct answer. Figure 2 summarizes the results.

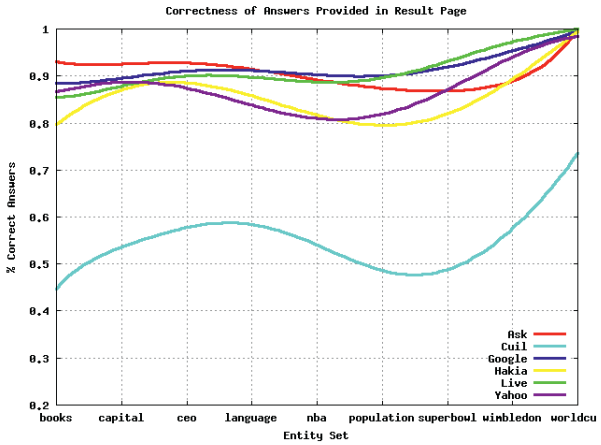

Figure 2. Correctness of answers provided in results pages

All the search engines (with the exception of Cuil) have performed surprisingly well during this invariance test. Most of the search engines seem to be able to provide the correct answer in the SERP for about 90% of the rephrases tested.

The great variety of URLs returned for the rephrases shows a lack of true understanding of the query, however, the use of popular subjects (as the World Cup or the Super Bowl) for which there is redundancy of information on the web and plenty of user generated content, helps simple keyword matching to capture and highlight at least one copy of the answer for each form of the query.

Given the richness of content of today's SERP and the often limited amount of space available in the browser, for a complete and accurate analysis it is important to determine where the answer is found in the result pages. After

having removed all the HTML tags and non-content (e.g. JavaScripts or CSS styles) from each page, we computed the position of the answer as a fraction of the total length of the page. Figure 3 averages the positions of the answers across all the ORA schemata used in this experiment:

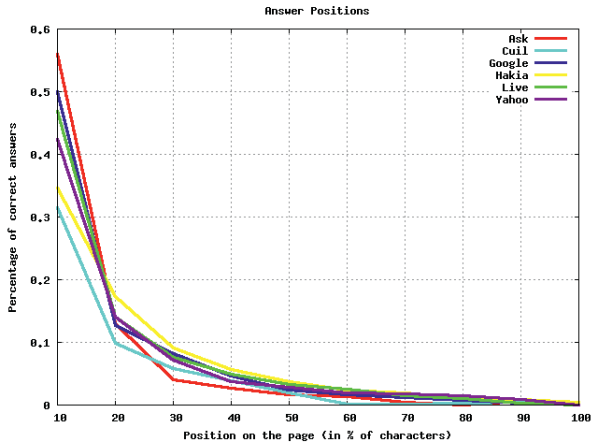

Figure 3. Average Position of Answers in result pages

The X-axis subdivides the position of the answer with respect to the total length of the page while the Y-axis indicates the fraction of queries which fell into that space. For example, from the graph above it is possible to observe that Ask.com provides the answer to 55% of the ORA queries in top 10% characters of each page, while Google only about 50% of the times. Surprisingly, Hakia provides only the 35% of its answers in the first percentile.

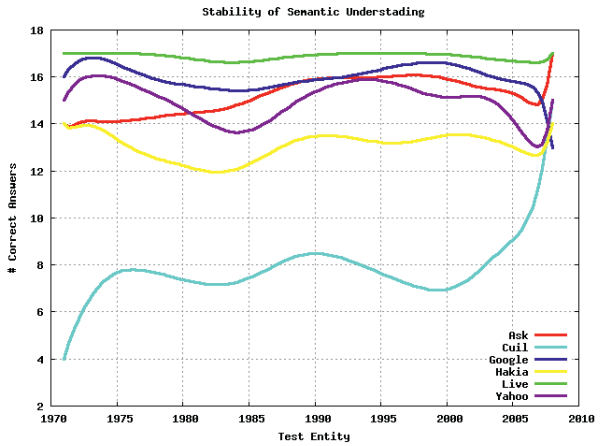

Figure 4. Stability of Semantic Understanding for Super Bowl dataset

Finally, we performed a stability test on each scheme to measure the stability of its answers across the various parameters. Figure 4 shows the fraction of rephrases which returned the correct answer for equivalent queries of "Super Bowl *year* Winner", across different years.

As previously said, in a truly semantic search engine the correctness of the answers provided by each rephrase should definitively not depend on the parameter chosen, given that the engine has such knowledge in its database. The results of this last experiment indicate Live as the most stable search engine with respect to the parameter, and Cuil as the least stable engine with an high sensitivity to the parameter used in the scheme. Given that the previous experiments confirmed that Live is not semantic, the high stability of results could once more be attributed to the popularity of the subject and the redundancy of content and tags around the topic.

Top-K Results Overlap The goal of this second part of the experiment is to measure the difference among the results returned by a search engine to different forms of the same query. An ideal semantic search engine should always return the same set of URLs all the times, achieving 100% overlap among the top N URLs returned for each rephrases. Unfortunately, Figure 5 demonstrated that this is not the case:

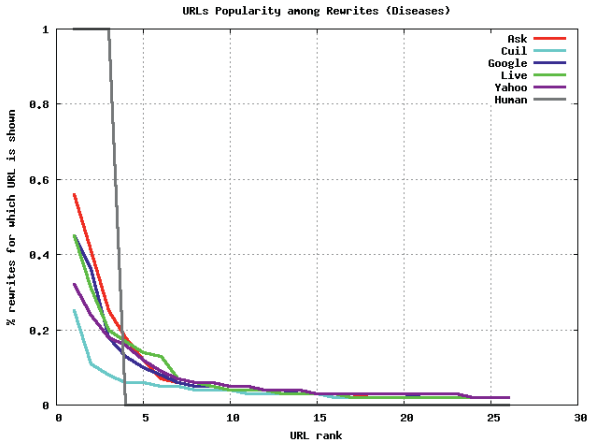

Figure 5. Stabiliy of top-3 results for Disease dataset

The graph of Figure 5 has been computed considering only the top 3 URLs shown for every query of the Diseases dataset (179 rephrases, 60 diseases). Every point in the graph represents the average fraction of rephrases for which the n-th most popular URLs has been shown to the user by each search engine. The plot shows, for example, how the most common URL for Ask.com is generally shown for about 55% of the rephrases and that the 3rd most popular URL for Live is generally shown for only 20% of the queries in the dataset. The gray line depicted in the graph represent an ideal Human Search Engine, which due to its perfect understanding of the queries always return the same top 3 results for every reformulation of the original query producing the sharp function depicted in the graph.

Entropy An ideal semantic search engine should be invariant to rephrasing and always return the same set of URLs, in the same order, for equivalent queries. This is especially true for the first URL which is assumed to be the most relevant to the user query.

In this last experiment we computed the entropy of each search engine, that is, the average degree of uncertainty of the first URL returned as result of equivalent queries.

During our test we used the rephrases of the ORA dataset (population, language, superbowl, books, wimbledon, CEO, president, NBA, worldcup, capital) and computed the probability of each top URL u returned for dataset x as $p(u) = \#first(u)/\#rewrites(x)$. Using these probabilities we calculated the entropy of each dataset d as $Entropy(d) = -\sum_{i=0}^{n} p(u_i) * log(p(u_i))$.

	Ask	Cuil	Google	Hakia	Live	Yahoo
pop	3.17	3.91	0.81	3.7	2.6	0.75
lan	3.71	5.26	0.96	4.51	3.33	0.84
sbw	2.13	3.61	2.07	3.37	1.89	3.17
bok	0.66	0.99	0.67	1	0.77	0.9
wbl	3.15	3.2	3.14	3.05	2.57	3.08
ceo	1.36	2.18	0.75	1.56	1.28	1.33
prs	0.56	0.98	0.32	0.82	0.43	0.33
nba	2.24	3.35	2.03	3.12	2.43	3.1
wcp	0.74	3.16	0.7	2.68	0.85	2.55
cap	1.19	1.61	1.11	1.69	1.15	1.36
avg	1.89	2.82	1.26	2.55	1.73	1.74
%	46.3	69.1	30.9	62.5	42.4	42.6

Table III
ENTROPY OF TOP-URL FOR ORA DATASET

Table III summarizes the results of the experiment. For each engine we reported the entropy obtained on each data set, as well as the average entropy among all the dataset, and its percentage value with respect to the worst case scenario (i.e. each rewrites returns a different set of URLs as top result). The smaller the entropy, the more semantic search engine is.

Considering the tested data sets, Google emerged as clear winner ahead of Live, Yahoo and Ask. The start-ups Hakia and Cuil showed results worse than expected. In the population set (pop) Yahoo slightly beated Google, but both resulted well ahead of the other engines..

These results are very preliminary and more massive tests are perhaps necessary before drawing ultimate conclusions, but they definitively give an idea of the current status of "semantic search".

While tackling more detailed experiments, it could be very interesting to measure Entropy at lower ordinals or consider top-K results as a set (i.e. ignoring ordering and considering the engine "stable" if the same set of results is returned for different rephrases of the same query).

IV. CONCLUSIONS

The results returned by a truly semantic search engine must be invariant for semantically equivalent queries, but even well-known engines currently fail to satisfy this condition. In this paper we proposed 3 types of invariance: URLs Stability, Binary Overlap Invariance and ORA Invariance for queries with "one right answer". We have extracted and generated different rephrasings for multiple query schemata, tested simple synonym replacements, and measured the effects of over-specification of the query when a redundant category term was added. AOL's query logs and Wikipedia-based lists of entities have been used to cons truct the queries used in our experiments. The experimental results summarized in the paper suggest the following:

1) Invariance of results for general queries is still poor. Today's search engines are very sensitive to the way queries are phrased. They are all mostly keyword based and far away from simulating human query understanding.
2) Queries with One-Right-Answer (ORA) seems well served by search engines, which manage to return the correct answers in their result pages with surprising indifference to the form of the query..

Unfortunately, (2) is mostly a consequence of the massive redundancy of information on the web and the recent increase of user generated content. Often, the answer to different equivalent forms of a question can be found in the search result page because:

1) There are multiple pages talking about the same "fact" in different ways. For topics of massive interest (e.g. the world cup) many people creates pages with same or similar content. The subtle differences on the language and structure used to present those information help search engines to find at least one copy of the information through simple keyword matching.
2) Manual or semi-automatic tagging has been done to the page, enriching the unstructured text originally provided with additional information which helps the search engine to better find and index its content.

The recent wide-spread adoption of Search Engine Optimization (SEO) techniques play also an important role in this problem. While those techniques are unfortunately often associated with Spam, their original intent was legitimate: help search engines to do a better job while indexing and ranking page contents and URLs.

A good example of this is Amazon.com, which includes in the description of its products some additional information (e.g. the author for books) which greatly increase the probability of being found and thus provides an answer to the user. Another example is Wiki.Answer.com, which carefully crafts the URLs of its pages to make sure they are easy to find (using a standard XML sitemap) and contain the

question answered in the page (for better keyword matching and ranking).

Thanks to these artifices, search engines are often able to provide answers to popular ORA queries without actually understanding them. The search engines are not semantic, but with the implicit "help" of content providers (which work hard to have their pages ranked first) and user generated content (e.g. Yahoo!Answers) appear to be so for those queries. But while this might work for popular queries, rare topics do not always get the same attentions and the stability of their answers decreases rapidly.

The instability of current ORA answers shows even among some of the "popular" results (e.g. the winner of the Super Bowl, Figure 4) presented in this paper: different objects of the same class enable different subsets of schemata to provide the answer. Ideally, they should all trigger the same amount of rephrases when the answer is known to the search engine, or not at all when it is not.

This is a negative consequence of relying too much on "external help" as oppose to trying to understand the user queries. A truly semantic search engine would take care of invariance at query level, clustering together all its possible rephrases into one unique concept to which answer, and would allow to deal equally well with popular and unpopular topics (from the world cup to the local soccer league).

The data collected also confirms that the stability of results under rephrasing for general queries is still poor in all the search engines. Our experiments with simple numeric synonym replacements (e.g. "10" with "ten"), as well as the ones which involved the addition of redundant category terms, indicate the heavy reliance on text matching. The keywords used in the queries, and their position, strongly influence the distribution and the order of the results returned. This is not acceptable in a semantic search engines, which should lift from the users' shoulder the burden of correctly optimizing its query and take care of it on its own.

A. Acknowledgments

We thank the Department of Computer Science at University of Iowa for providing access to computing resources. We also would like to thank Antonio Gulli, Yufei Pan and all our friends and collegues at Ask.com for their helpful reviews and comments.

REFERENCES

[1] *Physical Review*, 1957, vol. 106, no. 4, ch. Information Theory and Statistical Mechanics.

[2] W3C, *Semantic Web Activity*. [Online]. Available: http://www.w3.org/2001/sw/

[3] *Resource Description Framework (RDF) Model andSyntax Specification*, W3C. [Online]. Available: http://www.w3.org/TR/1999/REC-rdf-syntax-19990222/

[4] B.-L. Tim, J. Hendler, and O. Lassila, "The semantic web," in *Scientific American Magazine*, 2001.

[5] R. Guha, R. Mccool, and E. M, "Semantic search," pp. 700–709, 2003.

[6] S. Cohen, J. Mamou, Y. Kanza, and Y. Sagiv, "Xsearch: A semantic search engine for xml," in *In VLDB*, 2003, pp. 45–56.

[7] C. Rocha, "A hybrid approach for searching in the semantic web," 2004.

[8] S. B. Tm and F. Guerra, "Peer-to-peer paradigm for a semantic search engine," 2002.

[9] E. Brill, J. Lin, M. Banko, S. Dumais, and A. Ng, "Data-intensive question answering," in *Proceedings of the 10th Text REtrieval Conference*, 2001, pp. 393–400.

[10] S. Lawrence and C. Giles, "Context and page analysis for improved web search," *IEEE Internet Computing*, vol. 2, pp. 38–46, 1998.

[11] C. Kwok, O. Etzioni, and D. Weld, "Scaling question answering to the web," in *C.C.T. Kwok, O. Etzioni, D.S. Weld*, 2001, pp. 160–161.

[12] F. Y. Florence Duclaye and O. Collin, "Learning paraphrases to improve a question-answering system," in *11th Conference of the European Chapter of the Association for Computational Linguistics*, vol. Workshop in NLP for QA. Budapest, Hungary: EACL 2003, April 2003.

[13] A. Hedstrom, "Question categorization for a question answering system using a vector space model," Master's thesis, Uppsala University, Uppsala, Sweden, 2005.