The Internet Democracy: A Predictive Model Based on Web Text Mining

Scott Pion and Lutz Hamel

Abstract— **This thesis describes an algorithm that predicts events by mining Internet data. A number of specialized Internet search engine queries were designed in order to summarize results from relevant web pages. At the core of these queries was a set of algorithms that embodied the wisdom of crowds hypothesis. This hypothesis states that under the proper conditions the aggregated opinion of a large number of non-experts is more accurate than the opinion of a set of experts. Natural language processing techniques were used to summarize the opinions expressed on a large number of web pages. The specialized queries predicted event results at a statistically significant level. These data confirmed the hypothesis that the Internet can function as a wise crowd and can make accurate predictions of future events.**

I. Introduction

THIS paper describes a system that predicts future events by mining Internet data. Mining Internet data is difficult because of the large amount of data available. It is also difficult because there is no simple way to convert text into a form that computers can easily process. In the current paper a number of search engine queries were crafted and the results were counted in order to summarize the text of all of the web pages that are indexed by the Yahoo! search engine. At first glance it may seem unwise to include the opinions of all writers, as opposed to the opinions of experts only. The Internet is very open and anyone can write anything without having credentials. Wouldn't it be better to simply rely on a few web pages that are well respected? A recent book entitled *Wisdom of Crowds : Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations* [1] draws on decades of research in psychology and behavioral economics to suggest that experts often give inferior answers when compared to the averaged answers of a large crowd.

An excellent example of the accuracy of a large group occurs when one is trying to guess a quantity, such as an individual's weight or the number of jellybeans in a jar. One striking example from Surowiecki [1] was a contest to guess the weight of an ox. There were approximately 800 guesses, and a scientist computed the average of all of the guesses. The average was 1197 pounds, and the actual weight of the ox was 1198 pounds. This guess was better than any of the 800 individual guesses. This demonstrates the idea behind the wisdom of crowds hypothesis: The group as a whole can be very accurate even if no individual in the group is very accurate. The core idea is that some people will be too high, others too low, but in the end these biases will cancel out and an accurate measure will emerge.

Another obvious example of the wisdom of crowds is an open market. Many economists believe that open markets, such as stock or commodities markets, are so accurate that it is impossible to predict future prices. This is the well known "efficient market hypothesis" [2]. The efficient market hypothesis states that because all information is released to the public at the same time, market participants know what the value of any stock or commodity should be. Because the group knows what the current price should be, the asset cannot be overvalued or undervalued, so its future price cannot be determined. For example, when oil prices went up quickly and reached $70 a barrel during the year 2005, there were a number of people suggesting that oil would keep rising. The idea behind the efficient market hypothesis is that if everyone knows that oil will be worth $100 a barrel in 6 months, why would anyone sell it at $70 a barrel presently?

The idea of markets predicting events has become so common that some web sites refer to themselves as prediction markets [3]. One can buy and sell contracts on events such as who will win elections, who will win reality television programs, whether the Vice President of the U.S. will resign, and whether a case of bird flu will be found in the U.S. One of the most famous prediction markets is the Iowa Election Market [4]. Since 1988 the Iowa Election Market has been more accurate than traditional polling [5]. In the current study the TradeSports.com prediction market [6] predicted the November 2006 U.S. Senate, House, and gubernatorial elections with an accuracy of 93%.

The wisdom of crowds hypothesis has received a great deal of attention within computer science recently. Examples of mass participation include Google, Wikipedia, Youtube, MySpace, Facebook, prediction markets, and the Internet in general. Voting is a common way for a crowd's opinion to be aggregated. Google uses a voting procedure to rank which web pages are the most relevant to a person's Internet search query [7]. The pages that appear at the top

This work was supported in part by the University of Rhode Island.

Scott Pion is with the Department of Computer Science and Statistics, University of Rhode Island, Kingston, RI 02881 USA (phone: 401-874-2701; e-mail: pions@cs.uri.edu).

Lutz Hamel is with the Department of Computer Science and Statistics, University of Rhode Island, Kingston, RI 02881 USA (phone: 401-874-2701; e-mail: lutz@inductive-reasoning.com).

of a Google search are the ones that match the text of the user's query and have the highest page rank compared to the other matches. Page rank is determined by how many web pages link to a given web page. Also, if a page with a high rank links to another web page, this incoming link is weighted more heavily. In a sense, links to a page are counted like votes for a page. If many pages have links to a target web page, it is assumed that the target web page is relevant to an individual's search query. Google is using the wisdom of the entire Internet in order to rank the most relevant web pages.

It is important to note that a crowd is not always more accurate than an expert. Specific conditions must be present [1]. If a great deal of expertise is required then the expert may outperform the crowd. For example, if a decision about the result of a complex physics experiment were required, an expert may outperform a large crowd. In a chess match, a world champion would probably beat a random crowd of 1000 people that voted on every move. A crowd tends to be most wise when it is similar to a random sample of a population. In statistics the idea of the random sample is that if one randomly selects people from a population, one should get a diverse, representative group. When a crowd is making a decision, in order to avoid bias, diversity of opinion is very important. Each person should use some private information, even if it is only their personal interpretation of publicly known facts. Another factor that tends to make the crowd wise is independence. If individuals' opinions are determined by people around them, then the crowd may simply represent the opinion of the most persuasive member.

The basic measure used to summarize web pages in the current study is counting results from Internet search engines. Counting Internet search results has received little attention from the computer science community. Most research has involved studying the relationship between merit and the number of results returned by a Google search [8], [9]. Bagrow and his coauthors studied the relationship between the number of publications a scientist has produced and the number of search results that were returned by Google. A total of 449 scientists were randomly chosen from the fields of condensed matter and statistical physics. The searches took the form of: "Author's name" AND "condensed matter" OR "statistical physics" OR "statistical mechanics". The relationship between the number of search results and the number of publications in an electronic archive was found to be linear with an R squared of approximately 0.53. This result implies that there is a relationship between the number of publications and the number of results retrieved from an Internet search engine.

II. METHODOLOGY

A. *Hypotheses and Goals*

The goal of this project is to apply the wisdom of crowds hypothesis to the Internet. The hypothesis is that results from Internet search queries will correlate with the predictions of an open market at a level significantly greater than zero. The wisdom of crowds hypothesis is often applied to three specific types of predictions. These predictions are economic indicators, sporting events, and elections. We will attempt to predict events from these areas in this paper. Algorithms based on computational linguistics will be used to produce counts summarizing predictions. These counts will then be correlated with the actual results. For example, if most web pages express the opinion that the New York Yankees will win the World Series, then the New York Yankees should win.

The general methodology is to try to predict the outcome of events by counting and integrating results from a series of Internet search engine queries. These search counts will be compared to a market prediction and the event result itself. The market predictions in this study come from the sports betting market, election prediction markets, and consensus economic numbers. Sample data is displayed in tables in later sections. The market prediction is often expressed in probabilities. For example, the counts could be compared to the sports betting market, which will assign a certain team a higher probability of winning an event such as the Super Bowl. The sports betting market, like most open markets, is assumed by many to be efficient [10]. Therefore the web count prediction is unlikely to outperform or even perform equally to any market, but may be expected to make similar predictions.

It is hypothesized that the web counts should perform better when predicting the market than the actual event because, according to the efficient market hypothesis, the market is supposed to take into account all of the information that is currently available and make the best prediction. Therefore, there is less variability in the market data than the actual results, making the market data more stable and predictable.

Because the algorithms used in this paper have a great deal of noise associated with them, the hypothesis is that the web count predictions will outperform a chance level prediction at a statistically significant level. An example of this noise is trying to predict whether Hillary Clinton will win the New York Senate seat in 2006. In a process that will be described later, one of the Internet search queries that will be used is "Clinton will win." This query could refer to Bill Clinton winning the 1996 presidential election or Hillary Clinton winning the 2008 presidential election, neither of which is the target. Even a more exact statement like "The Patriots will win the Super Bowl" could refer to the 2006 Super Bowl, even though the attempt is to predict the 2007 Super Bowl. Unfortunately, using more exact queries such as "will win the 2007 Super Bowl" gets only 829 results, whereas a more general query such as "will win the Super Bowl" gets 96,000 results. This small sample size makes it impractical to use the more specific version. Ideally the query would be general enough to have a large sample size but specific enough to express the correct predicate. Because more general queries were used it was expected that a great deal of false positives would be encountered. This led to the hypothesis that any predictions should be more accurate than a chance prediction but certainly not close to 100% accuracy.

In summary, the main hypothesis is that the correlations between the Internet counts and the market data, and the correlations between the Internet counts and the actual results, will be significantly greater than zero at the $p < .05$ level.

B. General Methodology

Web searches were performed with the Yahoo! search engine [11]. The Yahoo! search web services API was used along with the Java programming language in order to automate the search process [12]. One of the problems with counting Internet search results is that the dates of creation for most web pages are not available [13]. To deal with this problem, searches were also performed on the Yahoo! News website. Yahoo! News searches provide the exact date and time of the publication of each result [14]. The news searches gave results no more than one month old. It may be suggested that if the news dates are so accurate, then only the news results should be used. Unfortunately, the number of results from news searches is very low, so the general web search was used in order to be assured that the number of results would not often be zero.

Other details of the methodology used are specific to the area that is being predicted.

C. Sporting Events

Automating the data gathering for sporting events relied heavily on examples from the "question answering" literature [15]. In question answering one tries to program a computer to find an answer to a question in natural language, such as "Who was the first American in Space?" The field of question answering relies heavily on the broader field of natural language processing. One of the steps in question answering is determining the type of expected answer. For example, the question "Who was the first American in space?" should return a proper noun. Another facet of question answering is formulating the question into one or perhaps a number of queries that are submitted to some type of search engine. For example, the question "Who was the first American in space?" may create the query "was the first American in space." The words preceding this query are then appended, leading to answers such as "Sheppard was the first American in space."

The basic algorithm for predicting sporting events is given in Fig. 1 and described in the following paragraphs. The algorithm essentially has two parts. The first part is finding all of the potential winners. For example, if one wants an answer to the question "Who will win the Super Bowl," one expects the answers to each involve a team. The second part of the algorithm is to search to get a count for each potential winner that was found in the first part.

Algorithm:

```
searchQuery = "will win" + targetEvent
for counter = 1 to 200
    priorWords = three words prior to searchQuery
    newPhrase = priorWords + searchQuery
    parse newPhrase
    properNounArray[counter]=firstProperNoun(newPhrase)
end for
get all unique properNouns
for each uniqueProperNoun + searchQuery
    nounCountArray = count of web search results
end for
nounCountMax = maximum(nounCountArray)
for each nounCount
    if(nounCount < 1000 and nounCount <0.01 *
    nounCountMax)
        delete nounCount from nounCountArray
    end if
end for
result = nounCountArray
```

Fig. 1. Algorithm to predict sporting event

Drawing on the question answering literature, the search query used was of the form "will win *event*," such as "will win the Super Bowl." As with typical search engines, the Yahoo! search engine includes a web page title and a small paragraph relevant to the search query for each result. These titles and paragraphs were searched to find the query string. For example the search "will win the Super Bowl" includes:

- ONLINE EXCLUSIVE: In My Mind: Why Baltimore will win the Super Bowl ...

The Penn, a college media publication. ... ONLINE EXCLUSIVE: In My Mind: Why Baltimore will win the Super Bowl. Nate Albright ...
media.www.thepenn.org/media/storage/paper930/news/2006/11/17/... - 48k -

The text from the query was located in each result. The three words preceding the query text were then appended to the sentence and saved in a text file. In the above example this would produce: "Mind: Why Baltimore will win the Super Bowl." The Stanford Lexicalized Parser [16] was then called from Java and run on the sentence to try to find the first proper noun preceding "will win the Super Bowl." The first proper noun was then appended to the query. Using the current example this would produce "Baltimore will win the Super Bowl." Each of these proper-noun-headed queries was then searched in order to get a result count. The results that were used were only those that were at least one percent of the maximum count or greater than 1000. This filtering was done to eliminate any small counts that would not affect the final results. An example of this output is:

Eagles will win the super bowl 276
Seahawks will win the super bowl 119
Falcons will win the super bowl 122

There is a potential for noise in this data in that a statement like "The Patriots will win the Super Bowl" could refer to the 2006 Super Bowl although we were interested in predicting the 2007 Super Bowl.

The sporting events that were predicted were the World Series of professional baseball, the Super Bowl of professional football, and the Bowl Championship Series of college football. The World Series occurred in 2006 and the Bowl Championship Series and the Super Bowl occurred in 2007. The goal was to predict the outcome of the event itself as well as the probability that each team will win the event as determined by the betting market [17], [18], [19]. For the World Series and Super Bowl the actual results were determined either by when the team was eliminated from the playoffs, or, if they were not in the playoffs, their final standing during the regular season [20], [21]. The data for professional football was sampled twice, once three and a half months before the Super Bowl, and once one month before the Super Bowl. The data for the World Series was sampled three weeks before the event. The data for the Bowl Championship series was sampled three months before the event. For the Bowl Championship Series the actual results were indicated by the AP top 25 college teams ranking [22]. Any team not in the top 25 was assigned a rank of 26. Table I displays a sample of the data for predicting the team that will win the World Series of baseball.

TABLE I
SAMPLE DATA FOR THE WORLD SERIES WINNER

Team	Sports Betting Market Probability	Web	News	Actual Finishing Position
NY YANKEES	0.50	2580	2	5
NY METS	0.20	328	1	3
MIN TWINS	0.10	0	1	5
SD PADRES	0.09	0	0	5

One can clearly see the expected correlation between the web count and the sports betting market probability. The NY Yankees are clearly the betting market favorite to win the World Series, and they are also the team that is most often associated with the search query "will win the World Series."

D. Economic data

The economic quantities studied were taken from the Yahoo! Finance web page [23]. The data included quantities such as GDP, inflation, unemployment rate, home sales, and others. The task of this project is to predict whether a given economic quantity will rise or fall. It is impractical to make more exact predictions, such as "New home sales will be 231,000 on January 28, 2007." Almost no one would write such an exact opinion on a web page. The algorithm for predicting economic quantities is given in Fig. 2 and described in the following paragraphs.

Algorithm:

```
searchQuery = targetQuantity  + "will"
for counter = 1 to 200
   postWords =  six words after searchQuery
   newPhrase = searchQuery + postWords
   parse newPhrase
   verbArray[counter]= firstVerb(postWords)
end for
get all unique verbs
exclude all verbs that are not synonyms of rise or fall
for each searchQuery + uniqueVerb
   verbCountArray =  count of web search results
end for
verbCountMax = maximum(verbCountArray)
for each verbCount
   if(verbCount < 1000 and verbCount <0.01 *
   verbCountMax)
       delete verbCount from verbCountArray
   end if
end for
for all verbs synonymous with "rise"
   riseCount += verbCount
end for
for all verbs synonymous with "fall"
   fallCount += verbCount
end for
result = riseCount and fallCount
```

Fig. 2. Algorithm to predict economic quantity

This algorithm is similar to the earlier one for predicting sporting events. It is convenient that in the English language opinions about the future are often expressed in a very standard manner [24]. The form is usually: *noun* will *verb*. Examples would be "inflation will rise" or "the Patriots will win." Therefore, for the economic data, each search query began with the form "*quantity* will," such as "inflation will." The six words following the query, or the words up to the end of the sentence, were then appended, such as "inflation will continue to rise as the Federal." The top 200 queries were then saved to a text file and parsed using the Stanford Lexicalized Parser [16]. These queries were parsed in order to find the first verb that followed the word "will." Verbs were used because the goal of this project is to discover what the quantities will do, such as rise or fall, which are both verbs. Verbs that were not synonymous with rise or fall were excluded. Each of these unique verbs was then appended to the phrase "*quantity* will," creating predictions such as "inflation will drop." The web results were then counted for each of these queries.

The results that were included were only those that were at least one percent of the maximum count or greater than 1000. This was done so that queries only getting a small fraction of the results would not be considered.

If the number of *fall* results was greater than the number of rise results, the prediction was that the quantity would *fall*, and if the number of *rise* results was greater than the number of fall results, the prediction was that the quantity would *rise*. In the case of economic data there was a consensus number that was taken from the Yahoo! Finance website [23]. Therefore one of the tests was to compare the predictions of the web search results with that of the consensus numbers. The other test was to determine if the web search results predicted the actual rise or fall of the quantity.

The data was collected weekly from 9/23/2006 until 1/21/2007. The data from the week starting on 11/19/2006 was not collected because the authors were not available to collect it. Table II displays sample economic data.

TABLE II
SAMPLE DATA FOR ECONOMIC QUANTITIES

Quantity	Web Rise	Web Fall	News Rise	News Fall	Actual Change	Market Predicted Change
Building Permits	155	6	1	0	-103	-13
Business Inventory	44	4	0	0	0.006	0.006
Capacity Utilization	376	88	1	0	0.001	-0.001

E. The 2006 congressional and gubernatorial elections

We attempted to predict the results of all of the U.S. Senate races, all of the gubernatorial races, and all of the House of Representatives races considered "key races" by CNN [25]. We also attempted to predict all of the House of Representatives races in the states with the seven largest number of House seats: California, Texas, New York, Florida, Ohio, Pennsylvania, and Illinois. If CNN reported a candidate as running unopposed then the race was not included in the study. The data was taken from CNN websites [26], [27]. Two candidates were selected to be studied for each race. The two candidates chosen were the ones most likely to win according to prediction market data [6].

The first part of making election predictions was determining which phrases to use in order to determine that someone was expressing the fact that they believe a candidate would win. For example, in the case of Hillary Clinton, possible phrases could be "Clinton will win", "Clinton will win the seat", or "Hillary Clinton will win the Senate seat." More complex phrases are more likely to express the proper belief, but less likely to be used commonly. The procedure for creating the queries was to use close races to determine what expressions were used most commonly. The process started with the simplest search queries and then added more complexity at each level. For example:

Clinton will
Clinton will win
Clinton will win the Senate
Hilary Clinton will win the Senate seat

Names from the top 10 most competitive races were chosen with half being Republican and half Democrat [6]. To start, the name and the word "will" (such as "Clinton will") were used as search query phrases. The verbs from the top 200 results were retrieved. These verbs were manually inspected in order to determine which ones expressed the belief that the candidate would win. The verbs selected were win, beat, defeat, take, hold, keep, and retain. The "*name* will" phrase was appended with these verbs and searched again. The phrases and the six words following the phrases from the top 50 web search results were collected. An example would be "McCaskill will win office, but neither will support." These phrases were examined in order to determine if there were any phrases that expressed the belief that the candidate would win. The two final phrases that were selected were simply "*name* will win" and "*name* will beat." These phrases allow for a large number of false positives, but the hope was that there would be enough of a signal to be detected above the noise.

Because of the possibility of a number of false positives, the last name of candidate alone was searched. For example, "Johnson will win" should be expected to get a large result count simply because Johnson is such a common name. The total counts for "*name* will win" and "*name* will beat" were added, and then that number was divided by the count of the name alone. This division by name was intended to have a standardizing effect in cases when one candidate's name was much more common than another candidate's name. In the results section "Web/name" is the number of web results divided by the number of results when searching the candidate's name. Similarly, "News/name" is the number of news results divided by the number of results when searching the candidate's name. Table III displays some sample election data.

TABLE III
SAMPLE DATA FOR ELECTIONS

Candidates	Web/Name	News/ Name	Market Prob.	Actual Votes
Giffords	1.22E-04	0	79.5	114,263
Graf	2.50E-05	0	17.5	89,104

III. Results and Discusssion

Table IV displays the correlations between the news and web results and the outcomes of sporting, economic, and electoral events.

TABLE IV

PREDICTING EVENT RESULTS AND MARKET PROBABILITES

	Corr.	N	95% c.i. lower	95% c.i. upper
Sports Results				
Web	-0.38	119	-0.52	-0.21
News	-0.29	119	-0.45	-0.12
Market	-0.62	119	-0.72	-0.50
Sports Market				
Web	0.55	119	0.41	0.66
News	0.47	119	0.32	0.60
Economic Results				
Web	0.10	165	-0.05	0.25
News	0.02	165	-0.13	0.17
Market	0.39	146	0.25	0.52
Economic Market				
Web	-0.02	157	-0.18	0.14
News	0.07	157	-0.09	0.22
Election Results				
Web/Name	0.28	478	0.19	0.36
News/Name	0.15	158	0.00	0.30
Market	0.89	162	0.85	0.92
Election Market				
Web/Name	0.30	162	0.15	0.43
News/Name	0.27	80	0.06	0.46

The correlations for the sports results are negative because those with the highest counts should have the lowest position; for example first place is considered position number one. The idea behind these correlations is that if a great deal of web pages made a prediction, then the event should occur, and the market should assign a high probability to the event occurring. All of the results, except for the economic data, confirmed the primary hypothesis. The correlations for the sports and election data are significantly different from zero, because the confidence intervals do not include zero. Also as expected, the web count and news month count correlations are slightly higher for the betting market data than for the actual outcomes of the events, although not significantly higher. As seen in the rows labeled “market,” the correlations were highest between the market probabilities and the actual events.

It is difficult to determine why none of the correlations between the various web counts and economic data were significant. It is possible that this data is simply too dynamic, and the web does not update itself quickly enough to keep up. For example, if someone writes “inflation will rise,” the writer could be referring to the year 2005, even though the target inflation data is released on December 2006.

IV. CONCLUSION

The majority of evidence collected for this project indicates that the Internet can be used to make predictions that are more accurate than chance levels. For all but the economic data, the web search results and the news search results correlated significantly with the actual results and the market data. Also as predicted, the correlations tended to be higher (but not significantly higher) for the market data than for the actual data. This was hypothesized to be the case because the market data is more predictable than the actual data. The highest correlations were between the market predictions and the actual events, which is a confirmation of the wisdom of crowds hypothesis and the efficient market hypothesis. The general pattern of results was similar for the majority of the areas studied. This pattern was that the market predictions of the actual results were the most accurate, followed by the web and news predictions of the market data, followed by the web and news predictions of the actual results.

There is a great deal of future work that could be done in this research area. Future research could further automate and generalize the techniques used in this paper. More specific queries could be used, and more advanced computational linguistics techniques could eliminate some of the false positives and false negatives that were encountered in the searches. The techniques that were used could also be used in a more general manner, predicting the outcomes of a large number of different events.

Although this data has told us a great deal about how the Internet can be mined to make predictions, it tells us even more about the Internet’s reliability. Because the Internet appears to be able to operate as an efficient market and a wise crowd, it tells us that the Internet shares some of the traits of a wise crowd. First, it tells us is that the opinions on the Internet are diverse. Second, it tells us that the opinions on the Internet are independent of other opinions. Finally, and most importantly, it tells us that the Internet as a whole appears to contain accurate information and can predict future events.

REFERENCES

[1] J. Surowiecki, *Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations*. Westminster, MD: Doubleday Publishing, 2004.

[2] E. F. Fama, “Random Walks in Stock Market Prices,” *Financial Analysts Journal,* September/October, 1965.

[3] Intrade. (January 24, 2007). *Intrade Trading Exchange.* Available: https://www.intrade.com/v2/

[4] University of Iowa, Henry B. Tippie College of Business. (January 24, 2007). *Iowa Electronic Market.* Available: http://www.biz.uiowa.edu/iem/markets/

[5] Wikipedia. (January 24, 2007). *Prediction Market.* Available: http://en.wikipedia.org/wiki/Prediction_market

[6] TradeSports.com (October 10, 2006). Available: http://www.tradesports.com/aav2/trading/tradingHTML.jsp?eventSelect=coupon_32&evID=coupon_32&updateList=true&showExpired=false#

[7] S. Brin and L. Page. (January 24, 2007). *The Anatomy of a Large-Scale Hypertextual Web Search Engine.* Available: http://infolab.stanford.edu/~backrub/google.html

[8] J. P. Bagrow, H. D. Rozenfeld, E. M. Bollt and D. Ben-Avraham, "How famous is a scientist? —Famous to those who know us." *Europhys. Lett.*, 67(4), 511–516, 2004. Available: http://people.clarkson.edu/~bolltem/Papers/epl8312.pdf

[9] M.V. Simkin and V. P. Roychowdhury. (September 28, 2006). *Theory of Aces: Fame by chance or merit?* Available: http://www.citebase.org/cgibin/fulltext?format=application/pdf&identifier=oai:arXiv.org:cond-mat/0310049

[10] S. Debnath, D. M. Pennock, C. L. Giles, and S. Lawrence, "Information incorporation in online in-Game sports betting markets," *ACM Conference on Electronic Commerce,* 258-259, 2003.

[11] Yahoo! (December 16, 2006). Available: http://www.yahoo.com/

[12] Yahoo! (December 16, 2006). *Yahoo! search web services.* Available: http://developer.yahoo.com/search/

[13] G. Tyburski, (December 16, 2006). *It's Tough to Get a Good Date with a Search Engine.* Available: http://searchenginewatch.com/showPage.html?page=2160061

[14] Yahoo! News (December 16, 2006). Available: http://news.search.yahoo.com/news/search?fr=sfp&ei=UTF-8&p=test

[15] A. Gelbukh, (2006). Computational Linguistics and Intelligent Text Processing. Berlin:Springer.

[16] D. Klein, *Stanford Lexicalized Parser v1.5.1.*

[17] VegasInsider.com. (October 4, 2006). *MLB Future Book Odds at VegasInsider.com, the leader in Sportsbook and Gaming information - MLB Odds, MLB Futures.* Available: http://www.vegasinsider.com/mlb/odds/futures/

[18] VegasInsider.com. (October 4, 2006). *Sportsbook at VegasInsider.com: Online Sports Betting, Free Sports Picks, Las Vegas Odds, Adult Gambling.* Available: http://www.vegasinsider.com/u/futures/NFL_1336.cfm

[19] VegasInsider.com. (October 4, 2006). *Sportsbook at VegasInsider.com: Online Sports Betting, Free Sports Picks, Las Vegas Odds, Adult Gambling.* Available: http://www.vegasinsider.com/u/futures/FBC_1435.cfm

[20] Wikipedia. (February 1, 2007). *2006 NFL Season*. Available: http://en.wikipedia.org/wiki/2006_NFL_season

[21] Wikipedia. (February 1, 2007). *2006 Major League Baseball season.* Available: http://en.wikipedia.org/wiki/2006_Major_League_Baseball_season

[22] ESPN. (February 1, 2007). *ESPN.com - NCF - 2006 College Football Rankings - Week 17.* Available: http://sports.espn.go.com/ncf/rankingsindex

[23] Yahoo! Finance (January 24, 2006). *Economic Calendar: Financial Calendars – Yahoo! Finance.* Available: http://biz.yahoo.com/c/ec/200701.html

[24] Wikipedia. (January 31, 2007). *Future Tense.* Available: http://en.wikipedia.org/wiki/Future_tense

[25] CNN (December 21, 2006). *CNN.com – Elections 2006.* Available: http://www.cnn.com/ELECTION/2006/pages/results/house/

[26] CNN (December 21, 2006). *CNN.com – Elections 2006.* Available: http://www.cnn.com/ELECTION/2006/pages/results/Senate/

[27] CNN (December 21, 2006). *CNN.com – Elections 2006.* Available: http://www.cnn.com/ELECTION/2006/pages/results/governor/