By Michael Ettredge, John Gerdes, *and* Gilbert Karuga

Using Web-based Search Data to Predict Macroeconomic Statistics

Tracking common search terms used on the Web can produce accurate, useful statistics about the unemployment rate. We hope to extend this approach to other economic statistics.

This study investigates the potential of using data about Web searches to predict an important macroeconomic statistic, specifically the number of unemployed workers in the U.S. Our underlying assumption is that people reveal useful information about their needs, wants, interests, and concerns via their Internet behavior, and that terms submitted to search engines reflect this information. Research indicates that the percentage of Web site visitors who are referred by search engines increased from 67% in 2001 to 88% in 2004 [4], so this data potentially offers a rich and timely source of information. The study finds that Web-based search data is associated with future unemployment data over the 77-week study period. This very preliminary result suggests search-term data might be useful in predicting other important macroeconomic statistics.

A large proportion of job-related information gathering is conducted using the Internet [1, 10, 12]. Of the 54% of the U.S. population that uses the Internet, 16% engages in online job search activities [12]. There is evidence that unemployment duration has decreased among some Internet job seekers [9]. The Internet is credited with overcoming information bottlenecks in key areas of the labor market, affecting how worker-firm matches are made, how labor services are delivered, and how local markets shape demand [1].

Two kinds of Internet resources are available for job seekers—corporate job-posting sites and employment portals that serve as online employment agencies (such as Monster.com and Jobs.com). Access to either kind of employment resource requires job seekers first to locate its Web site, which is commonly done using search engines.

We attempt to determine whether the frequency of search terms likely used by people seeking employment could enable analysts to anticipate the content of forthcoming federal monthly unemployment reports. We maintain that individuals and businesses can better manage their economic affairs when important information is available on a continuous rather than periodic basis. This assertion is subject to a few caveats. First, acquiring this continuous information must be cost effective. Second, the benefit to society increases as the continuous information is more widely distributed. Third, the continuous information should not be too noisy if it is to be useful.

In our view, the value of search-term data hinges on the third condition. This study investigates whether search-term data is associated with future unemployment data. We do so by regressing official U.S. monthly unemployment data against Web-based job search data from preceding weeks. We also compare the explanatory power of (aggregated) daily Web search data with the explanatory power of (aggregated) weekly U.S. data on new unemployment insurance claims. Our goal in this regard is to calibrate the usefulness of the Web-based search data against alternative, official unemployment information that is timely, broad-based, and useful. Unemployment insurance claims data has been successfully used to forecast economic activity [5]. Finding that Web-based search data does have incremental explanatory power beyond weekly jobs claim data would be significant, and could lead to the use of such data in other contexts.

Panel A: Dependent Variable

Variable	Definition/(Data Source)
MUnemp_t	Number of unemployed for month *t*, seasonally adjusted. (Bureau of Labor Statistics monthly 'current population survey' of 60,000 households. The survey description and unemployment data are available at www.bls.gov/cps.)

Panel B: Explanatory Variables

Variables	Definition/(Data Source)
$\mathrm{SRate1}_t$ ST	Based on single day surge data, four-week average number of searches using the six job search terms over period ending one week prior to end of month *t*; normalized by dividing by the total number of searches during the same period. (WordTracker's Top 500 Keyword Report published by Rivergold Associates, LTD.)
$\mathrm{SRate1}_t$ LT	Based on 60 day long-term data, four-week average number of searches using the six job search terms over period ending one week prior to end of month *t*; normalized by dividing by the total number of searches during the same period. (WordTracker's Top 500 Keyword Report published by Rivergold Associates, LTD.)
$\mathrm{Claim1}_t$ ST	A four-week total of seasonally adjusted, first-time unemployment insurance claims over period ending one week prior to end of month *t*. (Department of Labor weekly report, available at www. doleta.gov.)
$\mathrm{Claim1}_t$ LT	An eight-week total of seasonally adjusted, first-time unemployment insurance claims ending one week prior to end of month *t*. (Department of Labor weekly report, available at www. doleta.gov.)

Definitions are shown for the explanatory variables with one-week lead-times (that is, cutoff date for search term data is one week prior to end of month *t*). Variables with longer lead times are defined similarly.

Table 1. Summary of variables, data resources, and expected sign of association.

Source of Data for Dependent Variable

The Bureau of Labor Statistics (BLS) commissions a monthly current population survey of 60,000 U.S. households. The survey description and unemployment data is available at www.bls.gov/cps. We employed data obtained from the BLS Web site in August 2003. The BLS claims the Employment Situation Summary report provides the "most comprehensive measure of national employment and unemployment," and is the "primary source of data on employment status and characteristics of the labor force" (based on site content as of August 2003). BLS releases the unemployment report for each month during the first few days of the succeeding month. The report not only provides aggregate statistics, but also breaks down unemployment levels and rates by both age and gender. While earlier studies [2, 5, 6, 10] find differences in Internet usage by both age and gender, more recently there has been a trend toward gender equality in Internet usage [3, 10]. The report provides gender-specific data for two age groups: 16–19 years of age, and 20+.

Both the business and popular press commonly run stories based on the BLS reports. This coverage reflects a widespread and plausible belief among business professionals that monthly unemployment data is important enough to cause substantial movement in stock prices, and because the unemployment rate affects both wage rates and consumer spending.

Although the reports contain comprehensive information about the unemployment situation in the U.S., we focus on a key statistic: the monthly number of unemployed workers, which according to BLS is one of the most frequently requested statistics. (We repeated all tests using an alternative dependent variable: the monthly unemployment rate. Those results are not tabulated but are statistically similar to results obtained using monthly unemployment levels.) For each month *t* we designate the seasonally adjusted monthly number of unemployed as $\mathbf{MUnemp_t}$.

Sources of Data for Explanatory Variables

Search engine keyword usage data was extracted from WordTracker's Top 500 Keyword Report published by Rivergold Associates, Ltd. These weekly reports track keywords submitted to the Web's largest metasearch engines. WordTracker uses metasearch engines "because they give an unbiased view of searches. Millions of people check their Web site rankings daily in the major engines using software robots, and this can artificially inflate figures. This type of software is not generally used on metasearch engines" (based on WordTracker site content, Sept. 2003). Search engine positioning specialists, marketing personnel, and Web site managers use the WordTracker service to identify changing usage trends, so they can adapt their Web sites.

The reports contain two lists—the Top 300 Surge Report summarizing data over the previous 24 hours, and the Top 200 Long Term Keyword Report aggregating data over the previous 60 days. Both reports provide the most-used terms, and usage frequency, along with the number of searches during the period. This study covers the 77-week period from Sept. 15, 2001, to Mar. 1, 2003, which had 381 unique, heavily used terms in the Long Term report. (Ten reports were missing during this period. Data for those weeks was interpolated using data from preceding and succeeding weeks.) Using this list we identified six terms likely to be used by people seeking work, namely (in decreasing frequency of usage): job search, jobs, monster.com, resume, employment, and job listings. Monster.com describes itself as an online global career network with over 800,000 job listings.

	N	Min.	1^{st} Quartile	Median	3^{rd} Quartile	Max.	Mean	S. Dev.
MUnemp*	17	8,035	8,302	8,405	8,450	8,711	8,367	190,867
$SRate1_t$ ST	17	0.0208	0.0440	0.0465	0.0505	0.0623	0.0468	0.0099
$SRate1_t$ LT	17	0.0245	0.0287	0.0310	0.0365	0.0422	0.0324	0.0052
Claim1 ST*	17	1,552	1,576	1,623	1,666	1,961	1,651	108,686
Claim1 LT*	17	3,110	3,179	3,254	3,378	3,762	3,305	181,519

Variables are defined in Table 1.
* In thousands (that is, 8,035 = 8.035 million).
Descriptive statistics shown are for the explanatory variables with one-week lead-times (that is, cutoff date for search term data is one week prior to end of month t), but are representative of the variables with longer lead times.

Table 2. Descriptive statistics for study variables.

Weekly job search activity is found by summing the number of searches performed using these six job-search terms. To account for varying search activity, we normalize the search term usage data by dividing this sum by the total number of searches during the same period. Using surge data yields one-day rates. Four-week moving averages of these one-day rates (one per-weekly report) are calculated to obtain a short-term (ST) usage rate **$SRate_t$ ST**. Similarly, using the long-term data we compute a 60-day long-term (LT) usage rate **$SRate_t$ LT**.

We argue that information obtained using search terms will be more useful when it is available further ahead of the official report, and more strongly associated with official report data. Empirically, there should be a trade-off between these two criteria. We investigate this trade-off by employing variables with lead times varying from one to four weeks. This resulted in four short-term ($SRate1_t$ ST through $SRate4_t$ ST) and four long-term ($SRate1_t$ LT through $SRate4_t$ LT) variables for each month studied.

To evaluate their incremental usefulness, we require our search-term variables to compete against official weekly unemployment data. Newly unemployed persons file claims with state unemployment insurance agencies under the Federal-State Unemployment Insurance Program. This information is subsequently reported on the Department of Labor Web site (www.doleta.gov). As with the monthly BLS report, the business press often covers these weekly claims reports. We defined a short-term claim variable (**$Claim_t$ ST**) as a four-week moving average of the seasonally adjusted initial claims data. We also defined a long-term claim variable (**$Claim_t$ LT**) as an eight-week moving average of the seasonally adjusted initial claims data. As with the search term data, the weekly availability of the unemployment claims data allows the definition of eight variables with lead times varying from one to four weeks (**$Claim1_t$ ST, ..., $Claim4_t$ ST, $Claim1_t$ LT, ..., $Claim4_t$ LT**).

Table 1 summarizes our variable definitions, and Table 2 provides descriptive statistics. In our view, an important feature of the data is the limited variation of the dependent and explanatory variables during the time period studied. The monthly number of unemployed has a minimum of 8.035 million and a maximum of 8.711 million. (Due to article length requirements we do not tabulate descriptive statistics for all four versions of each explanatory variable. They are available from the authors.) The monthly unemployment rate (not tabulated) experienced even less variation. These narrow ranges are likely due in part to the seasonally adjusted nature of the data, but also might be indicative of fairly stable employment conditions during the period of the study.

Results

Given this study's experimental nature, and the limited number of months of search-term data, we do not attempt to use any time-series models, or to predict unemployment data in a hold-out time period. Instead, we employ all available observations of the monthly unemployment variable on the left-hand side of our equations, and investigate whether unemployment level is associated with search-term usage in preceding weeks. Establishing such a lead-lag relation is a first step in the prediction process. Initial results are presented in Table 3.

The explanatory variables in Panel A are the eight versions of the search-term usage rate. We find that two short-term search-term variables, **SRate1 ST** and **SRate2 ST**, have at best marginal significance, and the same is true of the associated model F-statistics. Variables **SRate3 ST** and **SRate4 ST** and their models were insignificant, and those results are not tabulated. ST variables also lacked explanatory power when included in models with LT variables (results not shown). In contrast, all four models including a long-term variable, **SRate1 LT** through **SRate4 LT**, are significant. Variables **SRate1 LT** and **SRate2 LT** lack significance when both are included in the same model (not tabulated). That model has similar explanatory power (R-square and F statistics) to the other models, which is an indication of multicollinearity. This is plausible since **SRate1 LT** and **SRate2 LT** largely employ the same data.

Panel A: Explanatory Variables = search-term usage rates

Explanatory Variables	Exp. Sign	Estimated Coefficient / p-value					
SRate1 ST	+	1.481 0.159					
SRate2 ST	+		1.752 0.100				
SRate1 LT	+			3.196 0.006			
SRate2 LT	+				3.161 0.006		
SRate3 LT	+					2.412 0.029	
SRate4 LT	+						2.503 0.024
No. of Observations		17	17	17	17	17	17
Adjusted-R^2		0.069	0.114	0.365	0.360	0.231	0.248
F-statistic p-value		2.193 0.159	3.068 0.100	10.216 0.006	9.990 0.006	5.816 0.029	6.266 0.024

Panel B: Explanatory Variables = search-term usage rates and unemployment claims

Explanatory Variables	Exp. Sign	Single-Variable Models				Two-Variable Models			
Claim1 ST	+	-1.971 0.067				-1.227 0.240			
Claim2 ST	+		-1.762 0.098				-0.898 0.384		
Claim1 LT	+			-1.815 0.089				-0.516 0.614	
Claim2 LT	+				-1.848 0.084				-0.540 0.598
SRate1 LT	+					2.589 0.021		2.380 0.032	
SRate2 LT	+						2.537 0.024		2.323 0.036
No. of Observations		17	17	17	17	17	17	17	
Adjusted-R^2		0.153	0.116	0.125	0.131	0.386	0.351	0.333	0.328
F-statistic p-value		3.884 0.067	3.104 0.098	3.296 0.089	3.415 0.084	6.032 0.013	5.334 0.019	4.991 0.023	

Table 3. Regression results. Dependent variable = monthly number of unemployed, seasonally adjusted.

We speculate the LT variables have more explanatory power than the ST variables because the sampling process used to generate the short-term job search data might introduce bias. The short-term values are an average of four weekly surge reports, with each report portraying activity that occurred on Friday of that week. It is plausible that search patterns might change throughout the week, necessitating more frequent sampling to obtain a representative measure of search activity. This potential sampling bias is not present with the long-term variables, since they integrate search activity over 60 days. Whatever the cause, the regression results discussed here suggest the LT search-term variables (but not the ST variables) are potentially useful in forecasting monthly unemployment rates.

The results in Table 3, Panel A, show that variables **SRate1 LT** and **SRate2 LT** have equivalent explanatory power. That is, the significance levels of the two LT variables, and the model adjusted R-squares, are similar in magnitude. We view the two-week lead time as superior from an information user's perspective since that date is further in advance of the monthly report, with no loss of explanatory power. Given this somewhat surprising outcome, we investigate the explanatory power of LT search-term rates computed with longer lead times of three weeks and four weeks. While the degrees of variable significance, and model adjusted R-squares, decrease with increased lead time, predictor variables **SRate3 LT** and **SRate4 LT** still have significant explanatory power, and their models are still significant.

Panel B in Table 3 addresses whether search-term data has incremental explanatory power beyond that provided by weekly federal unemployment claim activity. This comparison is interesting because the unemployment claims data is reported at frequent intervals, and thus provides an alternative to Web search data as a timely means of anticipating future monthly unemployment data. The unemployment claim data, unlike the monthly BLS current population survey, represents all jobless claims filed in the U.S., not just a sample.

Archival data going back to January 2000 is available for both unemployment data and new jobless claims. When data is pooled for all 41 available months, the association between new jobless claims (**Claim1 ST**) and the official unemployment rate is quite strong and positive, with an adjusted R-square of 0.939 (not tabulated). The model F-statistic also is large.

To compare the explanatory power of claims data to that of search-term data, it is necessary to rerun the analysis over the limited time period where search-term data is available. The model statistics in Panel B (F-statistics and adjusted R-Squares) indicate the explanatory power of job claims data is reduced over the study's shorter time period. The four jobs claims variables are significant in explaining the number of unemployed. However, in each case this explanatory power is lost when search-term variables are added to the model (see the two-variable models in Table 3, Panel B).

The single-variable regression results explaining the number of unemployed are interesting in that they indicate an inverse relationship between new unemployment claims and the number of unemployed people. This counterintuitive result might be specific to the 17-month period for which search-term data was available. Recall that the 41-month regression (not tabulated) shows a strong, positive relationship between these two variables. In any event, the claims variable loses its significant negative sign when the LT search variables are added to the models. This suggests the models might be better specified with the LT variables included, at least during the available sample period.

Our final analysis investigates whether the explanatory power of search-term data differs across age and gender groups. We compute individual monthly dependent variables for the number of unemployed, **MUnemp**, for different age and gender groups, using disaggregated BLS data. The available age categories are the 16–19 age group, and people aged 20 and over. We estimate regression models for each of the four age and gender combinations, using **SRate1 LT** and **SRate2 LT** as alternative explanatory variables (a total of eight regressions). The results (not tabulated) indicate that none of the regressions using as dependent variable the number of 16–19-year-olds unemployed has a significant F-statistic. In addition, none of the regressions using the number of unemployed females has a significant F-statistic. In contrast, both regressions for males aged 20 and above have significant F-statistics, and both **SRate1 LT** and **SRate2 LT** variables are significant.

This result should stimulate debate on Internet access and use by gender. Although the literature generally suggests equal Internet access by gender in the U.S., we find that males aged 20 and over appear more likely to use Internet search engines when seeking employment. It is possible that women use different job search terms than men, and we did not identify those terms in this study. We do not believe this to be the case. Other possibilities are that, compared to men, women believe that Internet job search is less likely to be effective for the positions they are seeking, or that women might prefer to rely on networks of friends even when seeking the same jobs as men.

Conclusion

In this article, we examined whether rates of employment-related searches by Internet users are associated with unemployment levels disclosed by the U.S. government in subsequent monthly reports. A positive, significant association is found between the job-search variables and the official unemployment data. We also observe an expected trade off between models' explanatory power and the length of the lead time between search-term usage and subsequent unemployment reports. Longer lead times are associated with lower explanatory power. We compare the explanatory power of the job search variables with variables computed using official weekly unemployment insurance claims data. Over the short available time horizon of this study, the significance of claims data was dominated by the explanatory power of search-term variables in joint models. We also investigated

We find that males aged 20 and over appear more likely to use Internet search engines when seeking employment. It is possible that women use different job search terms than men. We do not believe this to be the case.

THIS STUDY'S RESULTS SUGGEST THE VIABILITY OF USING SEARCH TERMS TO PREDICT OTHER IMPORTANT MACROECONOMIC DATA. WE CURRENTLY ARE INVESTIGATING THE ASSOCIATION BETWEEN SEARCH-TERM USAGE AND SUBSEQUENT MONTHLY REPORTS ON CONSUMER CONFIDENCE.

the relation between job-search term usage and unemployment data disaggregated by gender and age. Search-term usage is highly associated with unemployment data for males of age 20 and over. On the other hand, search-term usage is not associated with unemployment data for females of age 20 and above. This result suggests adult males are more likely to conduct Internet job searches than are their female counterparts.

Due to limitations on data availability, we do not actually predict hold-out unemployment data using Web search data. Rather, we undertake a preliminary investigation using the entire available data set to establish a statistical association between search-term usage and subsequent unemployment data. We are encouraged by the observed lead-lag associations, but speculate that careful additional work is required to define a role for search-term data in more extensive time-series prediction models. For instance, when we included prior month observations of the dependent variable (**MUnemp**) on the right-hand sides of our equations (for example, used period *t*-1 unemployment to explain period *t* unemployment), the LT search-term variables lost conventional significance (results not tabulated). This might reflect the data's low variability. Longer time series of job search terms should provide increased variability. Another limitation of the study is that some people using the Internet for job searches undoubtedly are employed, but seeking better jobs. This source of noise should bias our test results against what we nevertheless observe: a positive, significant association between search-term usage and lagged unemployment data.

Another avenue for future research on job searches would be to obtain data from large job site portals, such as monster.com, rather than metasearch engines. Further, this study's results suggest the viability of using search terms to predict other important macroeconomic data. We currently are investigating the association between search-term usage and subsequent monthly reports on consumer confidence. We hope readers will be stimulated to think of other associations.

References

1. Autor, D.H. Wiring the labor market. *J. Econ. Perspectives 15*, 1 (2000), 25–40.
2. Burrows, P., McWilliams, G., and Hoff R.D. PCs for everyone. *Business Week* (Mar. 23, 1998), 28–32.
3. Dholakia R.R., Dholakia N., and Kshetri N. Gender and Internet usage. *The Internet Encyclopedia.* H. Bidgoli, Ed. John Wiley and Sons, New York, 2003.
4. Elgin, B. Why the world's hottest tech company will struggle to keep its edge. *BusinessWeek* (May 3, 2004), 82–90.
5. Gavin, W. and Kliesen K. Unemployment insurance claims and economic activity. Federal Reserve Bank of St. Louis 84, 3 (May/June 2002).
6. Hafkin, N.J. Gender and ICT statistics. International Telecommunication Union, 3rd World Telecommunication/ICT Indicators Meeting. (Geneva, Jan. 15–17, 2003).
7. Jansen, B.J. and Pooch, U. Web user studies: A review and framework for future work. *J. Am. Soc. Inf. Sci. Technol. 52*, 3 (2000), 235–246; citeseer.nj.nec.com/cache/papers/cs/20657/http:zSzzSzjimjansen.tripod.comzSzacademiczSzpubszSzwus.pdf/a-review-of-Web.pdf.
8. Kuhn, P. and Skuterud, M. Job search methods: Internet versus traditional. *Monthly Labor Rev. 123* (Oct. 2000), 3–11.
9. Kuhn, P., and Skuterud, M. Internet job search and unemployment durations. *Am. Econ. Rev. 94,* 1 (Mar. 2004), 218–232.
10. McDougall, B. Cyber-recruitment—the rise of the E-labour market and its implications for the Federal Public Service. Labor Market Analysis Unit, Public Service Commission, Canada, Apr. 2001; www.hrma-agrh.gc.ca/research/labour-market/e-recruitment_e.pdf.
11. Srinivasan, K., Mukhopadhyay, T., Kraut, R., Kiesler, S., and Scherlis, W. Customer transaction behavior on the Internet: An empirical study. Workshop on Information Systems and Economics 1998; is-2.stern.nyu.edu/~wise98/pdf/five_a.pdf.
12. U.S. Department of Commerce. A nation online: How Americans are expanding their use of the Internet. Economics and Statistics Administration, National Telecommunications and Information Administration, Feb. 2002; www.ntia.doc.gov/ntiahome/dn/nationline_020502.htm.

Michael Ettredge (mettredge@bschool.wpo.ukans.edu) is a professor of accounting in the School of Business, University of Kansas, Lawrence.

John Gerdes (john.gerdes@ucr.edu) is an assistant professor of information systems in the Anderson Graduate School of Business, University of California, Riverside.

Gilbert Karunga (gkarunga@ku.edu) is an assistant professor in the accounting and information systems department, School of Business, University of Kansas, Lawrence.