Mining for Computing Skills

Andrew J. Aken

Department of Management, College of Business
Southern Illinois University
Carbondale, IL
ajaken@cba.siu.edu

Abstract

This research utilizes web content mining to retrieve job ads for graduates of Computing degree programs, extract the skills listed in those job ads, and analyze the data to determine which skills are most prevalent in the job market and which combinations of skills are most in demand.

Categories and Subject Descriptors K.3.2 [Computing Milieux]: Computers and Education – Computer and Information Science Education K.7.0 [Computing Milieux]: The Computing Profession

General Terms Measurement

Keywords curriculum, data mining, education, job market, skills, web content mining

1. Introduction

A crisis has been facing post-secondary schools offering Computing degrees for the past several years (Panko, 2008). Declining enrollments in these programs have had many schools looking into redesigning or eliminating them. The declining enrollment in these programs not only negatively impacts those colleges offering these programs, but it also has a dramatic impact on industry as well. Despite the current economic downturn and its impact on jobs throughout the US, Computing jobs are still growing at a significant pace and remain some of the fastest-growing jobs in the nation (Panko, 2008).

If the projections for growth in this sector are even remotely accurate, we will face a significant shortfall in qualified applicants for positions normally filled by Computing degree graduates. In Bill Gates' testimony before Congress earlier this year, he stated: "If the problem with high schools is one of quality, the issues at our universities is quantity. Our higher education system doesn't produce enough top scientists and engineers to meet the needs of the U.S. economy. According to the Bureau of Labor Statistics, we are adding over 100,000 new computer-related jobs each year. But only 15,000 students earned bachelor's degrees in computer science and engineering in 2006 and that number continues to drop." (Gates, 2008)

As one component of a proposed process for post-secondary schools to improve this situation, a comprehensive analysis of what skills industry expects of graduates of Computing degree programs is necessary. One method for collecting this information is through the analysis of job ads to determine what skills are specifically mentioned in the ads.

2. Data Collection

To collect the data for the analysis of the job postings on websites, the author developed a web content mining application which performed 2 primary functions:

1. Find all of the job postings for graduates in Computing degree programs and store the text of the job ads.
2. Extract and store the skills from the job ads.

To find the job postings, the author automated the task of performing a search on the Monster.com, HotJobs.com, and SimplyHired.com websites looking for references to the phrases synonymous with CS (e.g., "Computer Science" and "CS") which also had references to degree synonyms (e.g., Bachelors, BS, Masters, BA, Degree, etc.). Likewise, the software performed other searches for references to MIS (e.g., "Management Information Systems", "Management of Information Systems", "Computer Information Systems", MIS, and CIS) combined with the degree synonyms. Each search was also limited to jobs that had been posted in the last day so that the entire list of jobs returned from each unique search result would include less than 1000 postings (the limit to the number of jobs that can be retrieved from any of the job websites being used). The searches were automated to be run daily to retrieve the appropriate listings. Duplicated listings (based upon the website-specific job id) were excluded. The text of each of the web pages was subsequently stored in the database together with a version of the web pages with the HTML tags stripped out and another version with the word roots.

Prior research, analysis of job ads, and interviews were used to construct an initial collection of skill terms. Additional skill terms were extracted from the government's O*NET database and through proximity analysis of the skills found in the job ads. These skills included soft skills (e.g., leadership, oral communications, etc.), business skills (e.g., project management, supply chain management, etc.), technical skills (e.g., agile development, user interface design, etc.), and programming skills (e.g., c/c++, java, etc.). Skills which could also be used in common language were excluded from the search so that their importance would not be exaggerated (e.g., agile could refer to the software development methodology as well as a descriptive word for other types of work to be performed). The skills and synonyms of these skills (e.g., visual basic & vb) were also stored in a database in their word root forms. For each job page retrieved in the previous step, the posting was searched for the skills and their synonyms (e.g. visual basic, visual_basic, & visualbasic, vb, etc.). Information on the occurrences of these skills was also stored in the database.

The most frequently mentioned skills extracted from all Computing job ads and ads specifying Computer Science degrees are found in Figure 1. Subsequent cluster analysis of the skills found in each job ad was also performed to determine the combinations of skills being primarily looked for and to measure the relative quantity of job types (as determined by the cluster analysis). The results of these analysis and descriptions of future directions for the application can be found at www.dogs-it.org (the Degree-Oriented Guide to Skills in Information Technology).

OOPSLA'08, October 19–23, 2008, Nashville, Tennessee, USA.
ACM 978-1-60558-220-7/08/10.

Computer Science skills found through 2008-07-01 (495,926 job ads)

Skill Name	Occurrences	Frequency
Managing/Supervision	190,010	38.31%
SQL	144,051	29.05%
Testing	140,904	28.41%
Programming	140,316	28.29%
Security	131,131	26.44%
Software Development	129,545	26.12%
Quality	125,695	25.35%
Java/J2EE/J2P	114,691	23.13%
Microsoft operating systems	114,648	23.12%
HTML/XHTML/DHTML	107,596	21.70%
C/C++	106,506	21.48%
Financial	97,201	19.60%
Leadership	95,767	19.31%
Administration	95,202	19.20%
Customer Support	92,609	18.67%
UNIX operating systems	92,448	18.64%
Oracle databases	87,027	17.55%
Project Management	82,897	16.72%
Problem Solving Skills	79,283	15.99%
Business Strategy	77,953	15.72%
Software Engineering	76,336	15.39%
Analyst	74,369	15.00%
Certification	74,353	14.99%
Open-source operating systems	74,321	14.99%
Office Applications	73,926	14.91%
XML	72,852	14.69%
Generic databases	69,167	13.95%
Written Communications	67,360	13.58%
Microsoft databases	65,743	13.26%
Object Oriented Programming	60,912	12.28%
Troubleshooting	54,945	11.08%
.NET	54,809	11.05%
Initiative/Motivation to work	53,499	10.79%
Software Testing	53,470	10.78%
JavaScript	53,316	10.75%
SDLC	51,655	10.42%

Skills found for all Computing degree programs through 2008-07-01 (814,396 job ads)

Skill Name	Occurrences	Frequency
Managing/Supervision	330,597	40.69%
Security	216,195	26.61%
Financial	212,093	26.11%
Quality	200,663	24.70%
Testing	197,579	24.32%
Leadership	197,192	24.27%
SQL	184,887	22.76%
Business Strategy	179,041	22.04%
Programming	178,105	21.92%
HTML/XHTML/DHTML	170,322	20.97%
Administration	160,022	19.70%
Software Development	159,540	19.64%
Project Management	150,749	18.56%
Microsoft operating systems	149,703	18.43%
Java/J2EE/J2P	149,350	18.38%
Customer Support	138,963	17.11%
Certification	134,312	16.53%
C/C++	126,253	15.54%
Oracle databases	125,328	15.43%
Office Applications	124,565	15.33%
Analyst	123,749	15.23%
UNIX operating systems	117,418	14.45%
Problem Solving Skills	111,594	13.74%
Systems integration	109,518	13.48%
Written Communications	107,229	13.20%
Accounting	100,718	12.40%
Marketing	96,180	11.84%
Generic databases	92,515	11.39%
Open-source operating systems	91,673	11.28%
XML	88,403	10.88%
Initiative/Motivation to work	85,706	10.55%
Microsoft databases	85,429	10.52%
Software Engineering	84,643	10.42%
Business Process Design	83,735	10.31%

Figure 1: Skills found in 10% or more of the job ads

DOGS-it was designed to be the primary outlet for the vast quantity of information that has been collected through this continuously running web-content data mining application. The data being collected consists of information from job ads posted on Internet websites. The ads are parsed for various information, but primarily look for the skills required for the positions being advertised. The purpose of collecting this information is to try to determine the actual skills being sought by industry for graduates of particular degree programs.

Currently, this application is specifically retrieving job ads which require degrees in Computing programs. Other technically-oriented degree programs may be added in the future. As of 1 July 2008, the dogs-it application has retrieved, stored, and analyzed 814,396 job ads searching for 1,717 skills, and found 13,075,766 references to those skills in the job ads.

Once completed, this site will allow users to retrieve information regarding skill requirements for jobs in the selected degree programs. Information retrieved will vary depending upon the needs of the type of user and will be customized for:

- Students & JobSeekers
- Educators
- Industry
- Researchers

References

Gates, B. (2008, March 12). *Bill Gates: Testimony before the Committee on Science and Technology, U.S. House of Representatives.* Retrieved June 27, 2008, from Microsoft: http://www.microsoft.com/Presspass/exec/billg/speeches/2008/congress.mspx

Panko, R.R. (2008). IT employment prospects: beyond the dotcom bubble. European Journal of Information Systems 17, 182–197.