Spanish Keyword Search

(as used in Mi Próximo Paso)

Prepared for: National Center for O*NET Development
Post Office Box 27625
Raleigh, North Carolina 27611

Author: Jeremiah Morris
jm@whpress.com

Date: September 4, 2013

Updated February 18, 2014

Overview

Mi Próximo Paso (http://www.miproximopaso.org) is a career exploration website for Spanish-speaking job seekers and students, developed by the National Center for O*NET Development and sponsored by the U.S. Department of Labor. Mi Próximo Paso is modeled after the English-language site My Next Move.

A central feature of My Next Move is its keyword search, which suggests O*NET-SOC occupations based on a job title or descriptive phrase. Adapting the existing keyword search for a Spanish language website posed interesting challenges. This document outlines the search as implemented in Mi Próximo Paso. It also discusses some of the design decisions made, and testing methods used to ensure the accuracy and utility of the search results.

Weighted search features

The weighted keyword search used in My Next Move has several distinguishing features to aid career explorers:

- Key words are matched against a number of elements: occupation titles and descriptions, alternate titles, tasks, and detailed work activities. Matches on central elements like occupation titles have a greater influence on the results.
- Each word of a user's query contributes to the results, but words that are specific to only a few occupations are weighted more heavily.
- The search can match misspelled, truncated, or related words. For example, a search for "docter" will match "doctor," and a search for "nursing" will match "nurses."

The development of the English-language search is covered separately in *A Weighted O*NET Keyword Search* (http://www.onetcenter.org/reports/WWS.html). The steps in the Spanish-language search are covered in the section "The search algorithm" below.

Translation of occupation content

To provide relevant results in Spanish, it is important to have quality translations of the occupation content. We used a variety of sources for the translations:

- Occupation titles and descriptions were primarily based on the Spanish translation of the 2010 Standard Occupational Classification (SOC) taxonomy (http://www.bls.gov/soc/soc_2010_Spanish_Version.pdf), from the Bureau of Labor Statistics. Additional translations came from the Spanish version of the O*NET 4.0 Database and from translators at RTI.
- Lay titles, a set of over 40,000 job titles linked to O*NET-SOC occupations, were translated for this project by the Department of Labor.

- Tasks and detailed work activities were mostly machine translated using Google's translation services. Where available, existing translations from the Spanish version of the O*NET 4.0 Database and other sources were used.

The search database also supports multiple translations for a single piece of content. For example, a single lay title may have two translations loaded into the search: one from DOL's manual translation effort, and another from Google's machine translation. We found that this additional variety of translations gave a slight but measurable improvement in search results.

The weighted search implementations on the O*NET websites are kept up to date with the latest lay titles, task statements, Detailed Work Activities, and O*NET-SOC taxonomy information. While the algorithm itself changes infrequently, search results are improved through these data updates at least once per year. All new content is translated either by expert review or by machine. The Mi Próximo Paso search is updated simultaneously with the English language My Next Move.

Language-specific adjustments

In the English version of the keyword search, search terms and matching words are normalized by removing punctuation and changing all letters to lowercase. For the Spanish version, an additional step removes accents from characters: the letter "é" is transformed to "e" before searching. This allows matches to succeed even if the user cannot enter accented characters easily, perhaps from a smartphone or computer set up primarily for English speakers.

The rest of the algorithm remains the same for both languages, but three components are sensitive to the language in use:

1. **Spell check:** Our spell-check implementation uses GNU Aspell (http://aspell.net), for which English and Spanish dictionaries are available.
2. **Word stemming:** The Porter stemming algorithm for the Spanish language is used (http://snowball.tartarus.org/algorithms/spanish/stemmer.html). The English version of the search uses the Paice/Husk stemming algorithm, but a Spanish version of that stemmer was not available.
3. **Stopwords:** Words such as "and" or "the" are ignored to provide more relevant results. A different set of stopwords is used for Spanish text.

Evaluating the search results

The English version of the search was developed and refined over several years of regular use and testing. Whether these techniques would work well for a different language was a major concern. One approach would be to perform extensive testing with personnel fluent in Spanish and also familiar with the O*NET-SOC taxonomy, but that method requires significant testing resources, and relies heavily on subjective judgements. Instead, we used a blend of manual and automated evaluation, based on methods used previously to improve the English language search.

During the development of the English keyword search, a set of search queries were manually coded by analysts. The same set of queries was coded by the search algorithm, and the results were compared and scored. This allowed us to gauge whether an adjustment to the algorithm provided results closer to the gold standard of human coding.

For the Spanish search, we used a similar approach, but with the English search results forming the "gold standard" for comparison. We compiled a set of 275 queries, taken from the most popular search terms in My Next Move, and had them converted to Spanish by translators at RTI. We then compared the search results of the English and Spanish searches. Since the goal is to present the same content in both languages, a search for "police officer" on My Next Move should return many of the same occupations as a search for "oficial de policía" on Mi Próximo Paso. The top 20 and top 10 results from each search were used for comparison.

In the first evaluation, 62.2% of the top ten occupations returned by the English search were also found in the top top Spanish search results. After making adjustments to the spell checking and stopwords, and incorporating alternate translations for occupation content, this percentage increased to 63.6%. After the adjustments, the top occupation in the English search was also the top Spanish result 70% of the time, and found in the top three results 84% of the time.

Table 1 - Evaluation results

	Initial Implementation	Current Implementation
% of English results found in Spanish results (top 20)	62.65%	63.48%
% of English results found in Spanish results (top 10)	62.15%	63.63%
% where top English result is also top Spanish result	66.91%	70.18%
% where top English result is in top 3 Spanish results	81.09%	84.00%

Given that most queries are only one or two words long (making them very susceptible to differences in meaning across languages), the degree of correlation between the search results was encouraging. The automated results were followed up by several smaller tests with native Spanish speakers, who confirmed that search results seemed reasonable.

The search algorithm

From the initial user input, periods are removed, and any characters besides letters and numbers are treated as word separators. Accented characters are replaced with non-accented versions. All searching is case insensitive, so the normalized query can be seen as a list of words containing only lowercase letters or numbers.

All searchable content of the database is linked to an O*NET-SOC occupation; each occupation has a title, a description, and zero or more lay titles, tasks, and detailed work activities. Each of these elements is assigned to a thematic ring (see Table 2), which affects the relevance scoring.

For each unique word in the search query, a "word score" is calculated on a per-occupation basis as follows:

- The word is exactly matched against the content items of the database. The matches in each thematic ring for each occupation are tallied; the count of lay title matches is limited to 1, and the number of task or DWA matches is limited to 5. If the word matches a known stopword, all matches on the description, tasks, and DWAs are discarded. The final counts are then multiplied by the ring value (Table 2) and tier value (Table 3). For exact matches, the tier value is 4.
- The word is then stemmed, using the Porter stemming algorithm, and compared to the stems of the database content. The matches are tallied and multiplied as above; the stemmed tier value is 4.
- The word is matched as a substring; any word in the database beginning with the letters of the current word is considered a match. The matches are tallied and multiplied as above; the substring tier value is 2.

After these steps are applied, the implementation has a set of matching occupations and a word score for each occupation. The number of matching occupations is used to calculate a word frequency factor; this factor varies between 64 and 1. Table 4 shows the factors used. The word score is multiplied by the frequency factor to get a weighted occupation score for that word. The final raw score of an occupation is the sum of these weighted word scores.

If any words are misspelled, the process above is repeated for each unique spelling suggestion. The spelling suggestions have their own tier values, as shown in the chart below; note that the spelling substring tier value is zero, so that step may be skipped as it does not affect the score.

Once all words and spelling suggestions have been processed, two "exact match" phases are processed, where matching occupations are moved to the top of the results. The first phase checks against lay titles, and the second checks against occupation titles and singular equivalents, so that an occupation title exact match will be shown ahead of a lay title exact match.

In each "exact match" phase, the entire normalized query is checked against a normalized set of titles. If the query exactly matches, the score for the matching occupation(s) is adjusted as follows: the occupation's raw score is divided by 10, and the maximum raw score from the previous steps is added. Thus, each exact match's raw score is slightly higher than any previous score. When the scores are sorted, the exact matches will rank above all non-exact matches.

Search scores are not displayed in Mi Próximo Paso; they are only used to order the list of matching occupations. Occupations are displayed from largest to smallest score. A maximum of 20 occupations are returned.

Table 2 - Ring weights and filters

Thematic ring	Ring weight	Maximum count	Stopwords allowed
Occupation title	16	1	yes
Lay titles	16	1	yes
Description	8	1	no
Tasks	2	5	no
DWAs	1	5	no

Table 3 - Tier weights

Search tier	Tier weight
exact word	4
stemmed word	4
substring	2
exact word (spelling suggestion)	2
stemmed word (spelling suggestion)	2
substring (spelling suggestion)	0

Table 4 - Word frequency factors

Minimum number of matching occupations	Maximum number of matching occupations	Frequency factor
1	4	64
5	9	32
10	24	16
25	49	8
50	99	4
100	399	2
400	n/a	1

Summary

The keyword search is an important component of the My Next Move career exploration website, providing users a familiar interface for finding job information relevant to them. When building Mi Proximo Paso, special care was needed when adapting the keyword search for Spanish-speaking audiences. The language-specific changes involved query and data normalization, spell check, word stemming, and stopwords. The Mi Proximo Paso keyword search has become a key part of the O*NET Center's strategy to provide useful career information to new audiences.