can be attributed to the collective link text distribution on the web – many links with misspelled text point to sites with correct spellings. Such evidences can boost the confidence of a spelling correction model to suggest *vacuum* as a correction.

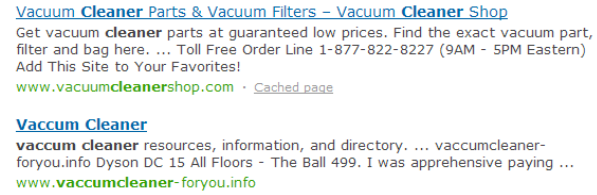

Figure 2. Sample search results for *vaccum cleaner*

The number of matched pages can be used to measure the popularity of a query on the web, which is similar to term frequencies occurring in query logs, but with broader coverage. Poor correction candidates can usually be verified by a smaller number of matched web pages.

Another observation is that the documents retrieved with correctly-spelled query and misspelled ones are similar to some extent in the view of term distribution. Both the web retrieval results of *vacuum* and *vaccum* contain terms such as *cleaner*, *pump*, *bag* or *systems*. We can take this similarity as an evidence to verify the spelling correction results.

4 Problem Statement

Given a query q, a spelling correction model is to find a query string c that maximizes the posterior probability of c given q within the confusion set of q. Formally we can write this as follows:

$$c^* = \mathbf{argmax}_{c \in C} \, Pr(c|q) \qquad \textbf{(1)}$$

where C is the confusion set of q. Each query string c in the confusion set is a correction candidate for q, which satisfies the constraint that the spelling similarity between c and q is within given threshold δ.

In this formulation, the error detection and correction are performed in a unified way. The query q itself always belongs to its confusion set C, and when the spelling correction model identifies a more probable query string c in C which is different from q, it claims a spelling error detected and makes a correction suggestion c.

There are two tasks in this framework. One is how to learn a statistical model to estimate the conditional probability $Pr(c|q)$, and the other is how to generate confusion set C of a given query q.

4.1 Maximum Entropy Model for Query Spelling Correction

We take a feature-based approach to model the posterior probability $Pr(c|q)$. Specifically we use the maximum entropy model (Berger et al., 1996) for this task:

$$Pr(c|q) = \frac{\exp(\sum_{i=1}^{N} \lambda_i f_i(c,q))}{\sum_c \exp(\sum_{i=1}^{N} \lambda_i f_i(c,q))} \qquad \textbf{(2)}$$

where $\sum_c \exp(\sum_{i=1}^{N} \lambda_i f_i(c,q))$ is the normalization factor; $f_i(c,q)$ is a feature function defined over query q and correction candidate c, while λ_i is the corresponding feature weight. λs can be optimized using the numerical optimization algorithms such as the Generalized Iterative Scaling (GIS) algorithm (Darroch and Ratcliff, 1972) by maximizing the posterior probability of the training set which contains a manually labeled set of query-truth pairs:

$$\lambda^* = \text{argmax} \sum_{c,q} \log Pr_\lambda(c|q) \qquad \textbf{(3)}$$

The advantage of maximum entropy model is that it provides a natural way and unified framework to integrate all available information sources. This property is well fit for our task in which we are using a wide variety of evidences based on lexicon, query log and web search results.

4.2 Correction Candidate Generation

Correction candidate generation for a query q can be decomposed into two phases. In the first phase, correction candidates are generated for each term in the query from a term-base extracted from query logs. This task can leverage conventional spelling correction methods such as generating candidates based on edit distance (Cucerzan and Brill, 2004) or phonetic similarity (Philips, 1990). Then the correction candidates of the entire query are generated by composing the correction candidates of each individual term. Let $q = w_1 \cdots w_n$, and the confusion set of w_i is C_{w_i}, then the confusion set of q is $C_{w_1} \otimes C_{w_2} \otimes \cdots \otimes C_{w_n}$ [1]. For example, for a query $q = w_1 w_2$, w_1 has candidates c_{11} and c_{12}, while w_2 has candidates c_{21} and c_{22}, then the confusion set C is $\{c_{11}c_{21}, c_{11}c_{22}, c_{12}c_{21}, c_{12}c_{22}\}$.

[1] For denotation simplicity, we do not cover compound and composition errors here.