Mining the Web
First Edition Errata

Soumen Chakrabarti

September 11, 2007

Thanks to Dalibor Fiala, Gordon Paynter, Horst Rothmeier, Marcin Sydow, and Dries Van Dyck.

1 Introduction

Page 2, second line from bottom, space missing: “humanity,and”.

Page 6, line 12: hyphen missing in “nonobvious” — should be “non-obvious”.

Page 7, line 7 from bottom: some people write “alumn” instead of “alum”.

2 Crawling the Web

Page 23, line 8: In the later version of the paper (page 14) they report an even better ratio: 70%/14%.

3 Web Search and Information Retrieval

Page 52, gamma code:

1. Unary code for $\lfloor \log x \rfloor$, which takes up $1 + \lfloor \log x \rfloor$ bits, followed by
2. $x - 2^{\lfloor \log x \rfloor}$ represented in binary, which takes up to another $1 + \lfloor \log x \rfloor$ bits.

Page 55, line 5 from bottom: Precision vs recall is more importantly a property of an IR system, not only a benchmark.

4 Similarity and Clustering

Page 85, in equation (4.1), $d_1 \neq d_2$ (consider each document pairs only once).

Page 86: The LHS of equation (4.4) should be $\langle p(\Gamma \cup \Delta), p(\Gamma \cup \Delta) \rangle$, not just $p(\Gamma \cup \Delta)$.

Page 86, lines 15–9 from bottom: The time complexity analysis is slightly unclear. Initially for each singleton group we form a Fibonacci heap of all other groups with the group-pair similarity as key. This takes $O(n^2)$ time overall. Then we have $n-1$ merge steps. In each merge step, to decide which pair to merge, we take the best score from n heaps, which takes $O(n)$ time. Now we actually perform the merge of Γ and Δ into a new group Φ. We compute Φ’s similarity to all other groups and form its heap in $O(n)$ time. We also inspect $n-2$ other heaps, delete occurrences of Γ and Δ from them (total time $O(n \log n)$) and insert a record for Φ in them (total time $O(n)$). Thus we get the total time as $O(n^2 \log n)$.

Page 89, line 13: $O(kn \log kn)$ is the same as $O(kn \log n)$, since k is $O(n)$ here.

Page 95: $0 < epsilon < 1$.

Page 106: In equation (4.27), all occurrences of α_i should be replaced by α_ℓ.

Page 109, Multiple cause mixture model: "The second half-step fixes $a_{d,t}$ and improves ..." should be "The second half-step fixes $a_{d,c}$ and improves ...".

Page 109 line 14: There is a mix of singular and plural here.

Page 110, second equation of (4.36): γ ranges over all the clusters.

Page 111, line 8 from bottom: should be "recall *and* precision".

5 Supervised Learning

Page 128 line 5: should be "... larger than *in* structured learning scenarios ..."

Page 132 line 5 from bottom: should be "... shown *in* Figure 5.2."

Page 134: equation (5.2) should read

$$\text{score}(c, d_q) = b_c + \sum_{d \in k\text{NN}(d_q)} s(d, d_q)[\![d \text{ is labeled } c]\!]$$

The last part, "$[\![d$ is labeled $c]\!]$", is missing.

Page 136 line 10 from bottom: missing 'n' in "unlimited".

Page 138 line 11 from bottom: missing brackets around "$C = 0$".

Page 140, just before equation (5.5), replace the rhs $\cdots = -\sum_{x,y} \log \frac{\Pr(x,y)}{\Pr(y)}$ with $\cdots = -\sum_{x,y} \Pr(x,y) \log \frac{\Pr(x,y)}{\Pr(y)}$.

Page 141 line 1 compared with text above equation (5.4) on page 139, explanation for the inconsistent comments about diverse document lengths: χ^2 uses the multivariate binary document model. Mutual information can deal with the multinomial document model (integer word counts), but does not perform length normalization. Fisher's discrimination can deal with length-normalized document vectors with real elements.

Page 144 line 6 from bottom, name collision/overloading: "T is multivariate term vector" vs 4 lines before T was declared to denote the set of terms. The understanding is that when we fix the vocabulary or feature set T, a vector of counts with one element for each element of T is also called T here.

Page 145 line 6: $T \cup C$ might look strange because they are from different domains, but at this point we are thinking of them as nodes in a Bayesian network.

Page 149, in equation (5.15), replace θ_t with $\theta_{c,t}$.

Page 150 equation (5.17) denominator and (5.18): some authors would write "dp" at the end of the integration by convention.

Page 151 line 6: "they have equal say" less clear than "they have equal influence".

Page 153 line 1 from bottom: the $\mathbf{x}$ on the lhs represents specific values of all elements of a vector; x in the rhs are elements of the vector $\mathbf{x}$.

Section 5.7.2 (Enhanced Parameter Estimation, bottom of page 155): Remove the sentence "I also introduced one technique toward better smoothing by exploiting document-length distributions."

Page 157: In step 6, Figure 5.9, the denominator should use $\theta^{\text{MLE}}_{c_j,t}$, not $\theta^{\text{MLE}}_{c_i,t}$.

The WH update rule in Equation (5.35) on page 164 does not discuss what to do with b. Here's a more complete description: Let each vector d be augmented by one extra element, always set to 1, and a corresponding extra dimension added to α, to simplify notation and get rid of b. The WH approach starts with some initial estimate $\alpha^{(0)}$ (with the extra dimension representing b), considers (d_i, c_i) one by one, and updates $\alpha^{(i-1)}$ to $\alpha^{(i)}$ as follows:

$$\alpha^{(i)} = \alpha^{(i-1)} - 2\eta(\alpha^{(i-1)} \cdot d_i - c_i)d_i.$$

Page 172 line 6 from bottom, line (c) 2: If we treat this definition literally (not as a shortening) constant literal "TRUE" would be the best. It seems to be better to use clear notion of positive and negative bindings here (as in the text) instead of "true" and "false".)

6 Semisupervised Learning

Page 182: In equation (6.3), the last summation in the denominator should also sum over τ (just like the first summation in the denominator).

Page 184 Figure 6.5 - there is no Bayes classifier measurements on the chart; however, it is mentioned in the caption. Details can be found in reference [168].

Page 192, Figure 6.8: In step 3 and step 7, replace w with v.

Page 192 line 4 from bottom: The word "small" before "gap" may mislead some readers since the interesting fact here is that there is an (observable) gap, not that it is small.

7 Social Network Analysis

Page 213, Figure 7.2, line 3 in while-loop: to be consistent, there should be arrows above h, as in $\vec{h}$; in line 4, a_0 and h_0 should be just $\vec{a}$ and $\vec{h}$.

Page 215, Figure 7.3, step 5: last factor should be $X(j)$ instead of $X(i)$.

Page 216: In equation (7.14) rhs, a_4 should be a_5.

Page 217, Figure 7.4, bottom right matrix is $E^\top E$.

Page 245, Figure 7.20, bottom graph, the labels read "in-degree" instead of "out-degree".

8 Resource Discovery

9 The Future of Web Mining

Bibliography

Reference [179] "suffic" should be "suffix".