Maximum Entropy Retrieval Using Iterative Proportional Fitting

Jung Jin Lee and Hyun Jo You

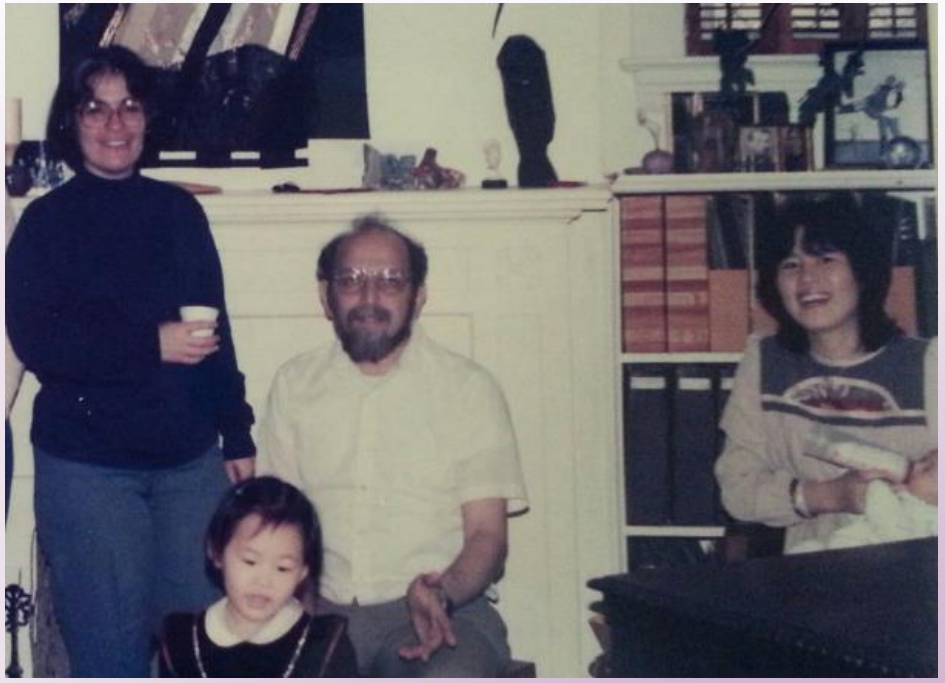
Soongsil University, Seoul, Korea

Overview

1. Introduction
2. Maximum Entropy Retrieval
3. Experiments
4. Discussions

1. Introduction

◎ *Paul Kantor & J.J. Lee (CWRU &Tantalus Inc. since 1982 ~)*



1. Introduction

◎ *Motivation*

- **Finding "relevant" documents from Scopus collections**
  - **Looking for a reviewer for the paper**

    'Nearest Neighbors Prediction Method for Mixed Logistic Model

➔ **too many (2,094,115) documents retrieved by title search**

➔ **retrieval ordering of documents using expert judgments**

Example :

Scopus Title Search :

‘Nearest Neighbors Prediction Method for Mixed Logistic Model

Nearest Neighbors
Entry published 2011 in SpringerReference

New prediction method for the mixed logistic model applied in a marketing problem
Journal Article published Oct 2013 in Computational Statistics & Data Analysis volume 66 on pages 202 to 216
Authors: Karin Ayumi Tamura, Viviana Giampaoli
Other IDs: S0167947313001369

Nearest Neighbors Problem
Entry published 2011 in SpringerReference

All-Nearest-Neighbors
Entry published 2011 in SpringerReference

A K -nearest neighbors survival probability prediction method
Journal Article published 7 May 2013 in Statistics in Medicine volume 32 issue 12 on pages 2062 to 2069
Authors: D.J. Lowsky, Y. Ding, D.K.K. Lee, C.E. McCulloch, L.F. Ross, J.R. Thistlethwaite, S.A. Zenios

kNNsim: k-Nearest neighbors similarity with genetic algorithm features optimization enhances the prediction of acti vity classes for small molecules
Journal Article published 29 Jul 2008 in Journal of Molecular Modeling volume 15 issue 6 on pages 591 to 596
Authors: Dariusz Plewczynski
Other IDs: 349

Secondary structure prediction with learning methods (nearest neighbors)
Chapter published 2009 in Scientific Computation on pages 66 to 81

2. Maximum Entropy Retrieval

by Lee & Kantor (JASIS 1991, 1995)

- **Ex. A set of documents, *R*, has two index terms A and B**

Boolean Components			Probabilty			Ordering by
No	term A	term B	NRel Docs	Rel Docs	Fraction	Rel Ratio
1	0	0	p(1,0)	p(1,1)	f1	q1
2	0	1	p(2,0)	p(2,1)	f2	q2
3	1	0	p(3,0)	p(3,1)	f3	q3
4	1	1	p(4,0)	p(4,1)	f4	q4
					1.0	

=> how to find p(i,j) given some expert judgments on A and B

◎ *Maximum Entropy Principle(MEP) Estimation*

$$\text{Maximize} \quad -\sum_{i=1}^{2^2}\sum_{j=0}^{1} p_{ij} \log p_{ij}$$

$$\text{subject to} \quad \sum_{j=0}^{1} p_{ij} = f_i \qquad i = 1, \ldots, 2^2$$

$$\frac{p(3,1) + p(4,1)}{f_3 + f_4} = V_A$$

$$\frac{p(2,1) + p(4,1)}{f_2 + f_4} = V_B$$

$$p(i,j) \geq 0$$

- If the number of terms are increased,
  => troubles to solve a nonlinear optimization

[Fenchel Duality Theorem]

If f is a closed proper concave function on R^k

and K is a closed convex cone in R^k, then

$$\sup_{\mathbf{x}\in K} f(\mathbf{x}) = -\sup_{\mathbf{y}\in K^*} f^*(\mathbf{y})$$

where $K^* = \left\{\mathbf{y} \in R^k;\ <\mathbf{x},\mathbf{y}> \leq 0 \quad \forall \mathbf{x} \in K\right\}$ is the dual cone

and $f^*(\mathbf{y}) = \inf_{\mathbf{x}} \left[<\mathbf{x},\mathbf{y}> - f(\mathbf{x})\right]$ is the concave conjugate of f

[Theorem] For a positive vector $\mathbf{n} = (n_1, \cdots, n_k)$ and a closed convex cone K containing the constant vectors,

$$\sup_{\log \mathbf{p} \in K} \prod_{i=1}^{k} p_i^{n_i} = \frac{1}{k} \exp\left[N\left(\inf_{\mathbf{p} \in \hat{\mathbf{p}} - K^*} I(\mathbf{p} \,||\, \mathbf{u}) \right) \right]$$

where $I(\mathbf{p} \,||\, \mathbf{u})$ is the Kullback - Liebler I divergence and $\mathbf{u} = (1/k, \cdots, 1/k)$

◎ *Iterative Proportional Fitting (IPF) : I-Projection*

[Theorem] Russhendorf(1995)

Let $\mathbf{m}_1, \mathbf{m}_2, \cdots, \mathbf{m}_T$ be the given marginals.

Starting with an initial measure $\mathbf{p}^{(0)}$, compute the *I*-projection $\mathbf{p}^{(1)}$ of $\mathbf{p}^{(0)}$ on the set of $\mathbf{m}_1$ i.e.,

$$I(\mathbf{p}^{(1)} \| \mathbf{p}^{(0)}) = \sum_{\mathbf{x} \in S} p_{\mathbf{x}}^{(1)} \log(p_{\mathbf{x}}^{(1)} / p_{\mathbf{x}}^{(0)})$$

Compute the *I*-projection $\mathbf{p}^{(2)}$ of $\mathbf{p}^{(1)}$ on the set of $\mathbf{m}_2$ and so on. i.e.,

$$\mathbf{p}^{(0)} \xrightarrow{\mathbf{m}_1} \mathbf{p}^{(1)} \xrightarrow{\mathbf{m}_2} \cdots \xrightarrow{\mathbf{m}_T} \mathbf{p}^{(T)} \xrightarrow{\mathbf{m}_1} \mathbf{p}^{(T+1)} \xrightarrow{\mathbf{m}_2} \cdots$$

Then the iteration converges to the unique probability measure $\mathbf{p}^*$ which satisfies all the given marginals.

◎ *IPF-procedure with MEP formulation*

Assume $\mathbf{p}^{(0)} = (p_1^{(0)}, \ldots, p_k^{(0)})$

Repeat

For $m = 1,2,\cdots M,1,2,\ldots,M,\ldots$

Minimize $\sum_{i=1}^{2^k} \sum_{j=0}^{1} p_{ij}^{(t)} \log \frac{p_{ij}^{(t)}}{p_{ij}^{(t-1)}}$

subject to

$$p_{i0}^{(t)} + p_{i1}^{(t)} = f_i, i = 1,\ldots,2^k$$

$$\sum_{x \in s_m} p_{x1}^{(t)} = V_m \sum_{x \in s_m} f_x$$

until $\mathbf{p}_{ij}^{(t)}$ converges.

◎ *Solution of IPF-procedure with Rules*

if $\mathbf{x} \in s_m$

$$p_{\mathbf{x}1}^{(t)} = V_m \frac{p_{\mathbf{x}1}^{(t-1)}}{\sum_{\mathbf{x} \in s_m} p_{\mathbf{x}1}^{(t-1)}} \qquad \text{if} \sum_{\mathbf{x} \in s_m} p_{\mathbf{x}1}^{(t-1)} \neq 0$$

if $\mathbf{x} \notin s_m$

$$p_{x1}^{(t)} = (1 - V_m) \frac{p_{\mathbf{x}1}^{(t-1)}}{\sum_{\mathbf{x} \notin s_m} p_{\mathbf{x}1}^{(t-1)}} \qquad \text{if} \sum_{\mathbf{x} \notin s_m} p_{\mathbf{x}1}^{(j-1)} \neq 0$$

3. Experiments

◎ LETOR3 OHSUMED test data

- LETOR
  - Microsoft Learning to Rank Datasets
  - LETOR 3.0 includes .Gov and OHSUMED
- OHSUMED
  - MEDLINE subset
  - 348,566 references
  - 106 queries
  - 16,140 query-documents pairs have been judge for relevance

◎ Queries

QueryID	QueryText
1	Are there adverse effects on lipids when progesterone is given with estrog en replacement therapy
2	pathophysiology and treatment of disseminated intravascular coagulation
3	anticardiolipin and lupus anticoagulants, pathophysiology, epidemiology, c omplications
4	reviews on subdurals in elderly
5	effectiveness of etidronate in treating hypercalcemia of malignancy
6	does estrogen replacement therapy cause breast cancer
7	changes seen of head mri after toxic exposure
8	work-up of hypertension in patient with horseshoe kidney

◎ Documents

Title and Abstract

.I 274229
.U
91000001
.S
Am J Emerg Med 9101; 8(5):373-8
.M
Abdominal Pain/ET; Adolescence; Adult; Aged; Aged, 80 and over; Appendicitis/CO/*RI/US; Child; Female; Human; Leukocytes/*; Middle Age; Predictive Value of Tests; Support, Non-U.S. Gov't; Technetium Tc 99m Aggregated Albumin/*DU.
.T
Evaluation of women with possible appendicitis using technetium-99m leukocyte scan.
.P
JOURNAL ARTICLE.
.W
The authors evaluated the use of technetium-99m albumin colloid white blood cell (TAC-WBC) scan in women with possible appendicitis. One hundred and nine women underwent 110 TAC-WBC scans. One woman had a second scan on a separate admission and was considered two individual patients in the analysis. Twenty-six women had appendicitis, 10 of whom had a perforated appendix at surgery. The TAC-WBC scan was indeterminate (abnormal but nondiagnostic for appendicitis) in 52 women (47%), nine of whom had appendicitis. Fifty-eight scans were read as positive or negative for appendiceal pathology. There were 16 true positives, 5 false positives, 36 true negatives, and 1 false negative. The predictive value of a positive scan was 76%, and the predictive value of a negative scan was 97%. The TAC-WBC scan was positive in 62% of patients with appendicitis and negative in 43% of the patients without appendicitis resulting in an overall accuracy of 47% in the 109 women. The main value of TAC-WBC scan in women with possible appendicitis is its high negative predictive value and the main problem with the TAC-WBC scan is its high indeterminate rate.
.A
Henneman PL; Marcus CS; Butler JA; Hall TA; Koci TM; Worthen N; Wilson SE.

◎ Query-Doc Relevance Judgments

query-1 : w1=advers, w2=effect, w3=lipid, w4=progesteron, w5=estrogen, w6=replac, w7=therapi

rel	docid	w1	w2	w3	w4	w5	w6	w7	numMatch
1	48192	0	1	1	1	1	1	1	6
1	256570	0	1	1	1	1	1	1	6
1	165240	0	1	1	1	1	1	1	6
1	335079	1	1	1	0	1	1	1	6
0	276804	1	1	1	0	1	1	1	6
0	244338	0	1	0	1	1	1	1	5
1	143821	0	1	1	0	1	1	1	5
0	285257	0	1	0	1	1	1	1	5
0	201684	1	1	0	0	1	1	1	5
1	111457	0	1	1	0	1	1	1	5

Experiment 1

P(rel doc) given each word are estimated with the relevance feedback of top 10 documents ranked by the number of matched terms.

q1 : w1=advers, w2=effect, w3=lipid, w4=progesteron, w5=estrogen, w6=replac, w7=therapi

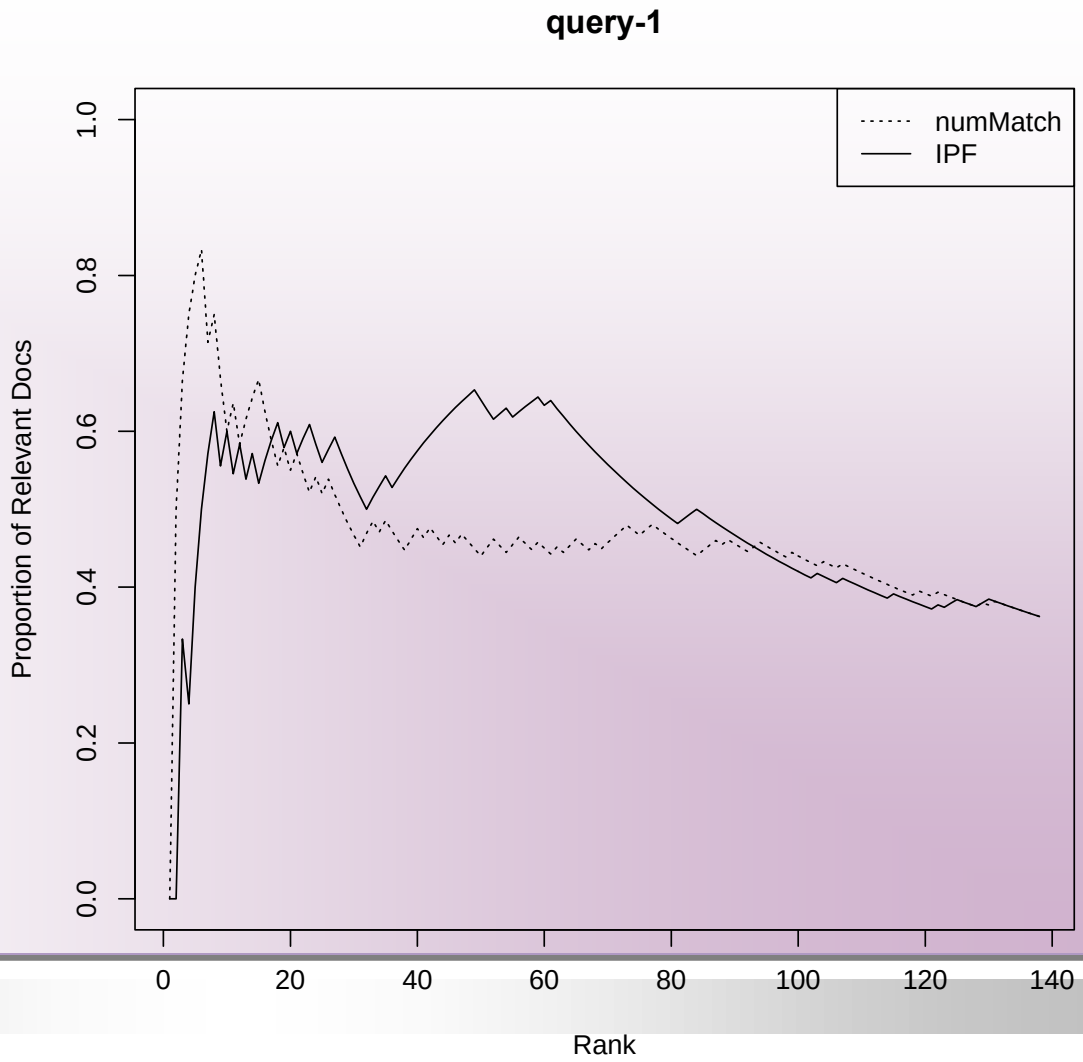
rel	docid	w1	w2	w3	w4	w5	w6	w7	numMatch
1	48192	0	1	1	1	1	1	1	6
1	256570	0	1	1	1	1	1	1	6
1	165240	0	1	1	1	1	1	1	6
1	335079	1	1	1	0	1	1	1	6
0	276804	1	1	1	0	1	1	1	6
0	244338	0	1	0	1	1	1	1	5
1	143821	0	1	1	0	1	1	1	5
0	285257	0	1	0	1	1	1	1	5
0	201684	1	1	0	0	1	1	1	5
1	111457	0	1	1	0	1	1	1	5

	w1	w2	w3	w4	w5	w6	w7
P(rel|w)	1/3	6/10	6/7	6/10	6/10	6/10	6/10

Experiment 1

IPF AveP = 0.547

numMatch AveP = 0.508



Experiment 2

If P(rel doc) for each word are provided by the user.

query-1 : w1=advers, w2=effect, w3=lipid, w4=progesteron, w5=estrogen, w6=replac, w7=therapi

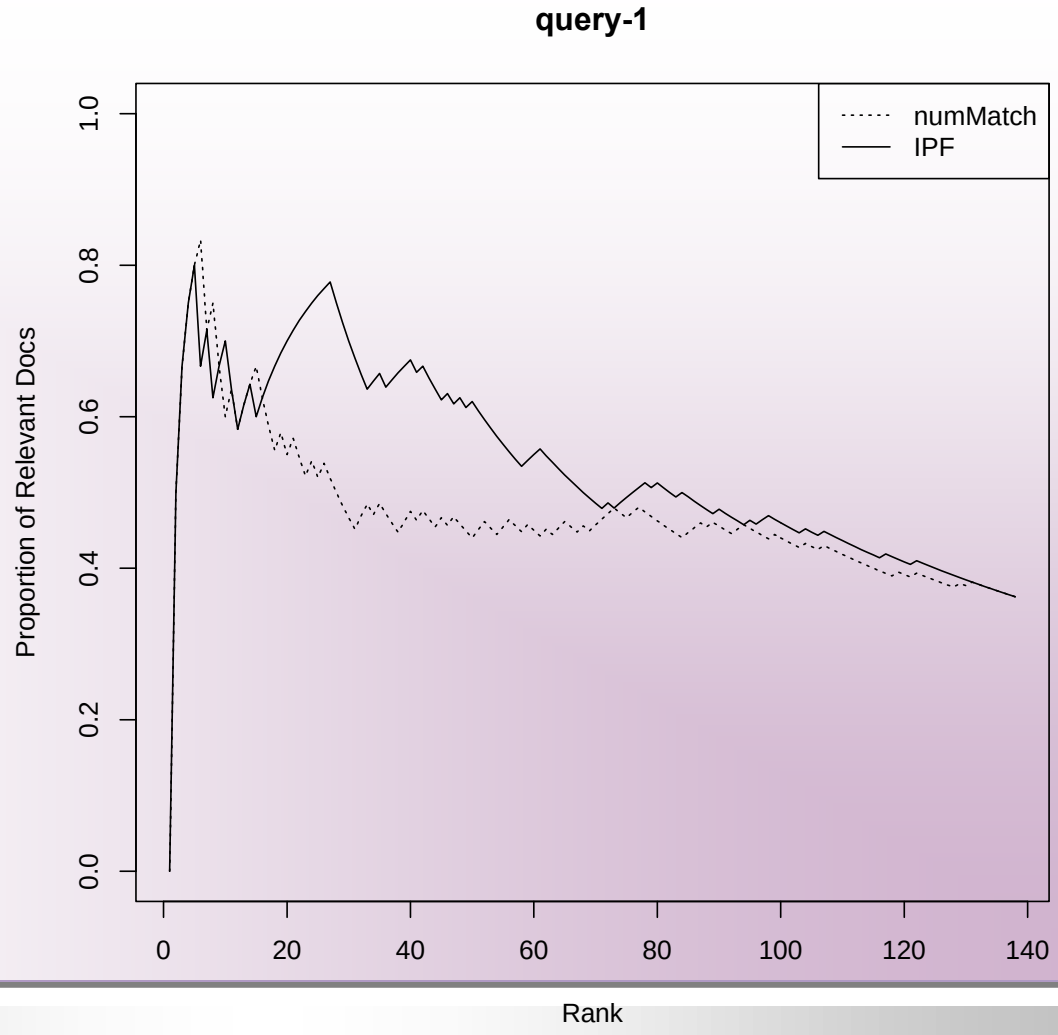
	w1	w2	w3	w4	w5	w6	w7
#rel_docs	9	46	33	7	39	35	41
#docs	32	103	54	38	110	92	100
P(rel|w)	**0.281**	**0.447**	**0.611**	**0.184**	**0.355**	**0.380**	**0.410**

Estimation with 10 documents:

	w1	w2	w3	w4	w5	w6	w7
P(rel|w)	**0.33**	**0.6**	**0.86**	**0.6**	**0.6**	**0.6**	**0.6**

Experiment 2

IPF AveP = 0.607
numMatch AveP = 0.508



Why Joint Probability Matters

Occurrence of a single term does not guarantee higher probability of relevance. It is possible that P(rel|-w) > P(rel|w) for a search keyword w.

q-1 : w1=advers, w2=effect, w3=lipid, w4=progesteron, w5=estrogen, w6=replac, w7=therapi

For documents without term wi:

	w1	w2	w3	w4	w5	w6	w7
#nonrel_doc	65	31	67	57	17	31	29
#rel_doc	41	4	17	43	11	15	9
P(rel|-w)	**0.38**	0.11	0.20	**0.43**	**0.39**	0.33	0.24

For document with term wi:

	w1	w2	w3	w4	w5	w6	w7
#nonrel_doc	23	57	21	31	71	57	59
#rel_doc	9	46	33	7	39	35	41
P(rel|w)	**0.28**	0.45	0.61	**0.18**	**0.35**	0.38	0.41

Ranking by IPF

Objective: higher ranking for documents with higher P(rel | w1,w2,…,wk)

rel :

- P(rel | w1,w2,…,wk)
- Proportion of relevant documents given match pattern

IPFrel:

- Joint probability P(rel, w1, w2, …, wk) estimated by IPF
- P(rel | w1, …, wk) = P(rel, w1,…,wk) / P(w1,…,wk)

Pattern	numMatch	IPFrel	rel
0110101	4	0.995	0.500
0111101	5	0.994	1.000
1110101	5	0.986	1.000
0110111	5	0.961	0.643
0110110	4	0.848	0.667
0110011	4	0.839	1.000
0111111	6	0.761	1.000
1110001	4	0.665	0.000
1110111	6	0.537	0.500
0100111	4	0.520	0.538
0101111	5	0.390	0.000
0110100	3	0.343	0.500
0110001	3	0.285	0.600
0111100	4	0.257	1.000
1100111	5	0.233	0.100
1110011	5	0.211	1.000
1100100	3	0.192	0.000
1110000	3	0.166	0.000
1110110	5	0.158	1.000
1101101	5	0.134	0.000

4. Discussions

- **If expert judgments and/or pilot judgments are close to actual relevance, MEP retrieval ordering shows some promising results.**
- **Effect of various types of expert judgment will be studied.**
- **Need more extensive study using real data.**

References

Dykstra R.L and Lemke, J.H., (1998) Duality of I Projections and Maximum Likelihood Estimation for Log-linear Models Under Cone Constraints, JASA, 83,No 402, 546-553.

Lee, J.J. and Kantor, P.B. (1991) A Study of Probabilistic Information Retrieval Systems in the Case of Inconsistent Expert Judgments. Journal of American Society for Information Science, V42, pp 166-172,

Kantor, P.B and Lee, J.J. (1998) Testing the Maximum Entropy Principle for Information Retrieval. Journal of American Society for Information Science, 49, No 6, 557-566.

Ruschendorf, L, (1995) Convergence of the Iterative Proportional Fitting Procedure. The Annals of Statistics, 23, No 4, 1160-1174.