ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ
ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ
ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΗΛΕΚΤΡΟΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ,
ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ & ΔΙΚΤΥΩΝ

Φοιτητής
Βερούλης Γεώργιος - (ΑΕΜ: 654) - geveroul@uth.gr

Περιεχόμενα

Υπόμνημα Εικόνων

Υπόμνημα Πινάκων

- **<u>Εισαγωγή</u>**

Στο πλαίσιο της διπλωματικής εργασίας, μου δόθηκαν οχτώ επιστημονικά άρθρα προς μελέτη τα οποία καλούνται να αντιμετωπίσουν τρείς γενικότερες κατηγορίες προβλημάτων. Προβλήματα δηλαδή που ασχολούνται με την ανάκτηση/αντιστοίχιση/ανάλυση οντοτήτων στο Web, την αντιστοίχιση/ανάλυση παραπομπών και την αποσαφήνιση ονομάτων, ειδικά για συγγραφείς paper. Θα γίνει αναφορά στην φύση αυτών των προβλημάτων και θα μελετηθούν οι μέθοδοι που προτείνονται για την αντιμετώπισή τους από τα επιστημονικά άρθρα.

1. Entity retrieval/matching/resolution στο Web

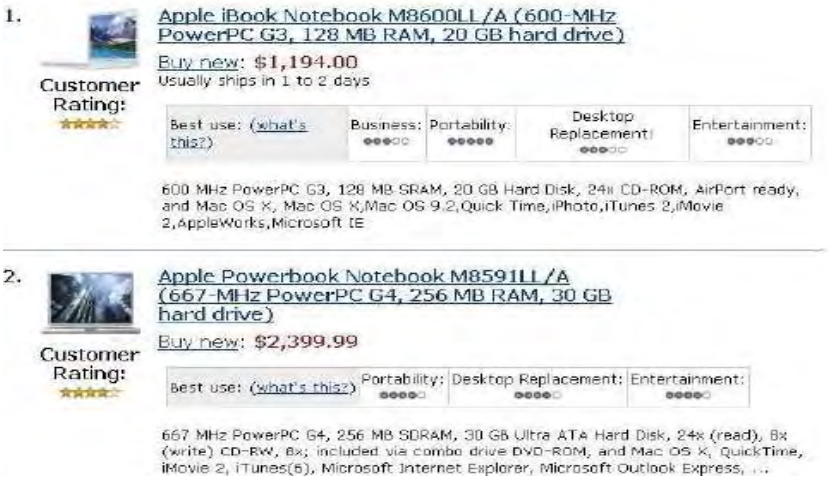
1.1 Introduction

Στις μέρες μας το Web αποτελεί την μεγαλύτερη και ευκολότερα προσβάσιμη πηγή πληροφοριών. Μεγάλη ποσότητα από αυτή την πληροφορία περιέχεται σε δομημένες μορφές αντικειμένων (data records) την οποία πολλές φορές ενδιαφερόμαστε να ανακτήσουμε. Καθίσταται επομένως αναγκαία η ανάπτυξη ενός αυτόματου αλγορίθμου ο οποίος θα μας επιστρέφει τέτοιου είδους πληροφορία την οποία εμείς με την σειρά μας θα μπορούμε να αξιοποιήσουμε με όποιο τρόπο θέλουμε.

1.2 Mining Data Records in Web Pages

Μας προτείνεται μία πλήρως αυτόματη μέθοδο με την οποία μπορούμε να εξορύξουμε όλα τα data records μιας ιστοσελίδας, η οποία σε σχέση με αντίστοιχες μεθόδους βελτιώνει την ποιότητα των επιστρεφόμενων αποτελεσμάτων. Ο αλγόριθμος που χρησιμοποιείται ονομάζεται MDR και βασίζεται σε δύο σημαντικές παρατηρήσεις:

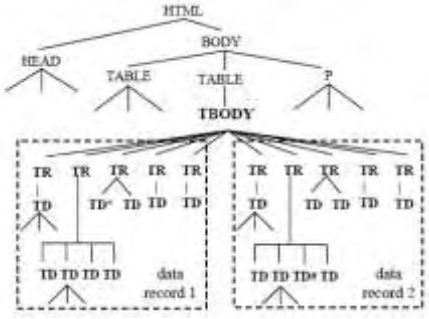
a) Ένα σύνολο από data records που περιγράφουν παρόμοια αντικείμενα βρίσκονται στην ίδια περιοχή μιας ιστοσελίδας και αναπαρίστανται με παρόμοια HTML μορφή (data regions). Θα μπορούσαμε να θεωρήσουμε τις HTML ετικέτες ως string και εκτελώντας συγκρίσεις να βρούμε τα αντικείμενα που αναφέρονται σε παρόμοια πράγματα, όμως το γεγονός ότι δεν γνωρίζουμε από πού αρχίζει και τελειώνει ένα data record όπως και κάθε data record μπορεί να περιέχει διαφορετικό αριθμό πληροφοριών, καθιστά αυτή την προσέγγιση απαγορευτική.
b) Οι HTML ετικέτες δημιουργούν μια σύνθετη δομή η οποία έχει μορφή δέντρου. Τα data records που περιέχουν παρόμοια μορφή πληροφορίας δημιουργούν ένα υπο-δένδρο με κοινό πατέρα-κόμβο.



Εικόνα 1: Κάθε υπολογιστής αποτελεί ένα data record με την περιοχή στην οποία βρίσκονται να ονομάζεται data region. Κάθε ετικέτα περιγράφει και από ένα χαρακτηριστικό του Η/Υ.

1.2.1 Τα βήματα της μεθόδου είναι τα εξής:

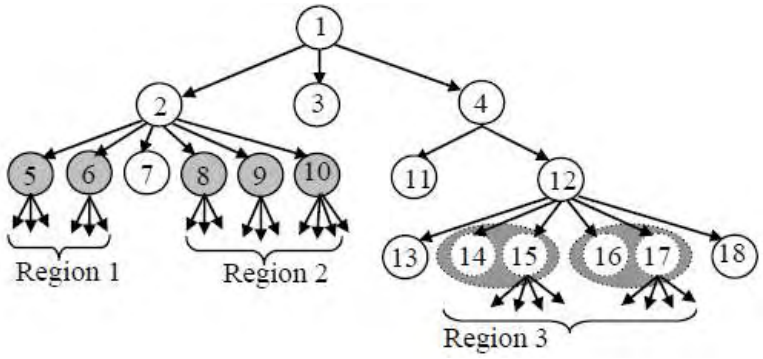
a) Δημιουργούμε το HTML tag tree μιας ιστοσελίδας όπου κάθε ζεύγος ετικετών (tags) του HTML κώδικα αποτελεί κόμβο του δέντρου. Στην HTML ένα ζεύγος ετικετών μπορεί να περιέχει τα data records που ψάχνουμε αλλά ίσως να περιέχει και άλλα φωλιασμένα tags. Πχ <title>Title of the document</title> μεταξύ των δύο ετικετών περιέχεται ο τίτλος ενός εγγράφου.



Εικόνα 2: Η μορφή ενός HTML tag tree

b) Κάνουμε εξόρυξη των όλων data regions μιας ιστοσελίδας που περιέχουν παρόμοιας μορφής data records. Ένας γενικευμένος κόμβος αποτελείται από ένα σύνολο γειτονικών κόμβων που έχουν τον ίδιο πατέρα κόμβο. Κάθε data region αποτελείται από μία συλλογή δύο ή περισσοτέρων γειτονικών γενικευμένων κόμβων. Στο παρακάτω δέντρο κάθε σκιασμένη περιοχή, αποτελεί γενικευμένο κόμβο. Οι γενικευμένοι κόμβοι 5,6,8,9 και 10 αποτελούνται από έναν απλό κόμβο άρα είναι μήκους 1 ενώ με την ίδια λογική οι

γενικευμένοι κόμβοι (14,15) και (16,17) είναι μήκους 2. Επιπλέον από την εικόνα 3 διακρίνεται ξεκάθαρα ποιοι από τους γενικευμένους κόμβους συνθέτουν μία data region.



Εικόνα 3:Παράδειγμα γενικευμένων κόμβων

1.2.2 Τα ερωτήματα που καλούμαστε τώρα να αντιμετωπίσουμε είναι από ποιον γενικευμένο κόμβο ξεκινάει μια data region και από πόσους κόμβους αποτελείται ένας γενικευμένος κόμβος σε κάθε data region.

a) Για να απαντήσουμε στο πρώτο ερώτημα απλά προσπαθούμε από κάθε κόμβο να ελέγξουμε αν ξεκινάει μία data region.

b) Για το δεύτερο ερώτημα εκτελούμε συγκρίσεις string με όλους τους δυνατούς συνδυασμούς γειτονικών κόμβων και με βάση τα αποτελέσματα τους αναγνωρίζουμε κάθε περιοχή. Για παράδειγμα σε περίπτωση ελέγχου τεσσάρων κόμβων (1,2,3,4) οι συγκρίσεις που θα γίνουν θα είναι μεταξύ των κόμβων με ετικέτες (1,2),(2,3),(3,4) και των συνδυασμένων κόμβων (1-2,3-4). Ο MDR αλγόριθμος που υπολογίζει κάθε δυνατό συνδυασμό σε κάθε κόμβο του tag tree είναι ο εξής:

Αλγόριθμος 1: MDR

```
Algorithm MDR (Node, K)
1. If TreeDepth(Node) ≥ 3 then
2.     CombComp(Node.Children, K);
3.     for each ChildNode ∈ Node.Children
4.         MDR(ChildNode, K);

CombComp(NodeList, K)
1. for(i = 1; i <= K; i++)
```

```
2.      for(j = i; j <= K; j++)
3.          if NodeList[i + 2 * j - 1] exists then
4.              St ← i;
5.              for(k = i + j; k < Size(NodeList); k + j)
6.                  if NodeList[k + j - 1] exists then
7.                      EditDist(NodeList[St....(k - 1)], NodeList[k....(k + 1 - 1)]);
8.                      St ← k + j;
```

Ειδικότερα ο MDR προσπελαύνει το tag tree από τη ρίζα προς τα φύλλα και σε κάθε εσωτερικό κόμβο, μέσω της CombComp εκτελεί συγκρίσεις string, βασισμένες στον normalized edit distance αλγόριθμο [23][24], σε διάφορους δυνατούς συνδυασμούς των παιδιών-δέντρων. Στη συνέχεια, αφού λάβουμε υπόψη τα αποτελέσματα των συγκρίσεων και ορισμένες παραμέτρους, βρίσκουμε τους υποψήφιους γενικευμένους κόμβους και data regions. Οι παράμετροι είναι οι εξής:

- αν μία data region καλύπτεται από μία άλλη τότε αναφέρεται μονάχα αυτή του υψηλότερου επιπέδου μαζί με τους γενικευμένους κόμβους που την αποτελούν.
- αν έχουμε έναν αριθμό παρόμοιων string τότε οποιοσδήποτε συνδυασμός μεταξύ τους θα είναι παρόμοιος. Συνεπώς εξετάζουμε οι γενικευμένοι κόμβοι που θα συνθέτουν μία data region να έχουν το ελάχιστο δυνατό μήκος.
- Αποτελεί ανάγκη ο ορισμός μιας ελάχιστης τιμής ομοιότητας που θα αποφασίζει αν δύο string είναι παρόμοια

Αφού βρεθούν οι υποψήφιοι γενικευμένοι κόμβοι και data regions, ο αλγόριθμος που παρουσιάζεται στην συνέχεια αναγνωρίζει ποιοι από αυτούς βρίσκονται σε μία ιστοσελίδα. Ως είσοδο δέχεται την τιμή κατωφλίου T, ένα κόμβο Node και μία μέγιστη τιμή κόμβων K ανά γενικευμένο κόμβο. Το Node.DRs περιλαμβάνει τις data regions κάτω από τον κόμβο Node ενώ στην tempDRs αποθηκεύονται προσωρινά οι data regions που παραμελήθηκαν από κάθε παιδί του Node.

Ο αλγόριθμος προσπελαύνει το δέντρο από την ρίζα προς τα φύλλα. Καθώς κατεβαίνει προσδιορίζει τις υποψήφιες data regions κάθε κόμβου ενώ κατά την επαναφορά του, πριν κατεβεί σε άλλο κλαδί, απορρίπτει αυτές που επικαλύπτονται από ένα γονέα data region του Node.DRs. Οι υπόλοιπες αποθηκεύονται στην tempDRs. Μετά το πέρας του αλγορίθμου το

σύνολο $Node.DRs \cup tempDRs$ θα περιέχει τις data regions του υποδέντρου με ρίζα τον Node . Στην συνέχεια αφού ληφθούν υπόψη οι συγκρίσεις που διεξήχθησαν στο προηγούμενο βήμα και η τιμή κατωφλίου που ορίσαμε, η αναδρομική διαδικασία IdentDRs βρίσκει τις data regions ορίζοντας ποιοι γενικευμένοι κόμβοι τις αποτελούν. Σε κάθε αναδρομή επιστρέφεται η επόμενη maxDR data region που καλύπτει το μέγιστο αριθμό παιδιά-κόμβων. Για κάθε ενδεχόμενο βρίσκεται η πρώτη data region αποτελούμενη από ένα σύνολο γενικευμένων κόμβων και στη συνέχεια ενημερώνεται ανάλογα η μέγιστη maxDR τιμή. Εξασφαλίζεται ότι θα ληφθούν υπόψη οι μικρότεροι γενικευμένοι κόμβοι, εκτός και αν οι μεγαλύτεροι καλύπτουν περισσότερους συνολικά κόμβους. Η τελευταία διαδικασία tempDiffDRs θεωρεί αυτές τις data regions πραγματικές και τις αποθηκεύει.

Αλγόριθμος 2: FindDRs Algorithm

Algorithm FindDRs(Node, K, T)

1. if $TreeDepth(Node) >= 3$ then
2. $Node.DRs \leftarrow IdentDRs(1, Node, K, T)$;
3. $tempDRs \leftarrow NULL$;
4. for each $Child \in Node.Children$ do
5. $FindDRs(Child, K, T)$;
6. $tempDRs \leftarrow tempDRs \cup UnCoveredDRs(Node, Child)$;
7. $Node.DRs \leftarrow Node.DRs \cup tempDRs$

Procedure IdentDRs(start, Node, K, T)

1. $\max DR = [0,0,0]$;
2. $for(i = 1; i <= K; i++)$
3. $for(f = start; f <= i; f++)$
4. $flag \leftarrow true$;

```
5.          for( j = f ; j < size(Node.Children); j + i)
6.              if Distance(Node, i, j) <= T then
7.                  if flag = true then
8.                      curDR[3] ← [i, j, 2 * j];      flag ← false;
9.                  else curDR[3] ← curDR[3] + i;
10.             else if flag = true then break;
11.         if (max DR[3] < curDR[3]) and
            (max DR[2] = 0 or (curDR[2] <= max DR[2]) then
12.             max DR ← curDR;
13.     if (max DR[3]!= 0) and
        (max DR[2] + max DR[3] − 1!= size(Node.Children)) then
14.         return {max DR} ∪ IdentDRs(max DR[2] + max DR[3], Node, K, T);
15.     return NULL;
```

```
Procedure UnCoveredDRs(Node, Child)
1.  tempDiffDRs ← NULL;
2.  for each data region DR in Child.DRs do
3.      if DR not covered by any region in Node.DRs then
4.          tempDiffDRs ← tempDiffDRs ∪ {DR};
5.  return tempDiffDRs;
```

c) Αναγνώριση των data records από κάθε data region. Για την επίτευξη κάτι τέτοιου λαμβάνουμε υπόψη τον περιορισμό πως αν ένα γενικευμένος κόμβος περιέχει δύο ή περισσότερα data records τότε αυτά θα πρέπει να αναφέρονται σε παρόμοια πράγματα. Τα

data records είναι κόμβοι και θα βρίσκονται στο ίδιο επίπεδο με τον γενικευμένο κόμβο ή ένα επίπεδο χαμηλότερα από αυτόν.

1.2.3 Πειραματικά Αποτελέσματα

Σημαντική αποτελεί η αποτελεσματικότητα του MDR αλγορίθμου έναντι των συστημάτων OMINI [25] και IEPAD [26] πράγμα που επαληθεύεται από παρακάτω αποτελέσματά. Πιο συγκεκριμένα, από τις 46 ιστοσελίδες που επιχειρήθηκε να γίνει εξόρυξη δεδομένων ο MDR παρουσιάζει 99.8% επιτυχής ανάκληση και 100% ακρίβεια, ποσοστά σαφώς πολύ ανώτερα σε σχέση με το 39% ανάκλησης 60% περίπου ακρίβειας των υπολοίπων δύο.

1.3 A Comparison of String Distance Metrics for Name-Matching Tasks

Πολλές φορές χρειάστηκε σε διάφορα είδη προβλημάτων να ληφθεί απόφαση για το ποσοστό ομοιότητας δύο string, πράγμα που κατέστησε αναγκαία την ανάπτυξη ανάλογων αλγορίθμων. Το paper αυτό μελετάει αλγορίθμους οι οποίοι μπορεί να προσεγγίζουν με διαφορετική λογική το πρόβλημα κάθε φορά. Στο τέλος συγκρίνονται μεταξύ τους με σκοπό να αποφασίσουμε ποιοι είχαν την καλύτερη απόδοση.

Μέθοδοι

1.3.1 Edit-distance like functions

Δέχονται ως είσοδο δύο strings και στην έξοδο μας δίνουν έναν πραγματικό αριθμό. Όσο μικρότερος είναι αυτός ο αριθμός τόσο περισσότερα κοινά στοιχεία μοιράζονται τα string. Το μέγεθος του δηλαδή εξαρτάται από το πλήθος των λογικών πράξεων που χρειάζονται για να μετατρέψουμε το πρώτο αλφαριθμητικό στο δεύτερο. Η μέθοδος του Levenshtein στηρίζεται σε αυτή τη λογική και προκειμένου να υπολογίσει την απόσταση μεταξύ δύο string a και b χρησιμοποιεί την παρακάτω σχέση:

$$lev_{a,b}(i,j) = \begin{cases} \max(i,j), if \quad \min(i,j) = 0 \\ \min \begin{cases} lev_{a,b}(i-1,j)+1 \\ lev_{a,b}(i,j-1)+1 \\ lev_{a,b}(i-1,j-1)+[a_i \neq b_j] \end{cases} \quad ,\text{otherwise} \end{cases}$$

Για παράδειγμα, η απόσταση Levenshtein ανάμεσα στις λέξεις "kitten" και "sitting" είναι 3, σύμφωνα με τις παρακάτω πράξεις:

a) **k**itten → **s**itten (substitution of "s" for "k")
b) sitt**e**n → sitt**i**n (substitution of "i" for "e")
c) sittin → sittin**g** (insertion of "g" at the end)

Γενικά συμπεραίνουμε πως η απόσταση Levenshtein:

- έχει πάντοτε ως ελάχιστη τιμή τη διαφορά μεγέθους των δύο string
- έχει ως μέγιστη δυνατή τιμή το μήκος του μεγαλύτερου string
- είναι μηδέν μόνο όταν τα δύο string είναι απόλυτα ίδια

Παρόμοια λογική έχει και η μέθοδος Monger-Elkan [30] με την διαφορά ότι οι τιμές της απόστασης έχουν εύρος [0,…,1].

Προτείνεται σαν καλή επιλογή επίσης η μέθοδος Jaro [31],[32] και οι παραλλαγές της οι οποίες βασίζονται στο πλήθος και το είδος των κοινών χαρακτήρων μεταξύ δύο αλφαριθμητικών μικρού μήκους όπως ονόματα συγγραφέων. Πιο συγκεκριμένα θεωρούμε δύο string $s = a_1...a_k$ και $t = b_1...b_L$ τα οποία μοιράζονται m κοινούς χαρακτήρες. Ορίζουμε στην συνέχεια τον όρο T ως το μισό πλήθος των μεταθέσεων προκειμένου να μετατρέψουμε το s στο t, για να καταλήξουμε στην τελική σχέση

$$Jaro(s,t) = \frac{1}{3}\left(\frac{m}{|s|} + \frac{m}{|t|} + \frac{m-T}{m}\right)$$

Μία αξιόλογη παραλλαγή της Jaro μεθόδου αναπτύχθηκε από τον Winkler [33] στην οποία λαμβάνεται υπόψη το μήκος P με τους περισσότερους κοινούς προθεματικούς χαρακτήρες των δύο string. Ορίζεται ως $P' = \min(P,4)$ για να καταλήξουμε στην σχέση

$$Jaro-Winkler(s,t) = Jaro(s,t) + \frac{P'}{10}(1 - Jaro(s,t))$$

Για παράδειγμα, αν μου δοθούν τα string *MARTHA* και *MARHTA* θα έχουμε:

- $m = 6$
- $|s| = 6$
- $|s| = 6$
- $T = \frac{2}{2} = 1$

Η απόσταση Jaro θα ισούται με

$$Jaro(s,t) = \frac{1}{3}\left(\frac{6}{6} + \frac{6}{6} = \frac{6-1}{6}\right) = 0.944$$

Ενώ για την Jaro-Winkler θα έχουμε

- $P' = 3$,επομένως

$$Jaro-Wienkler(s,t) = 0.9444 + \left(\frac{3}{10}(1-0.9444)\right) = 0.961$$

1.3.2 Token-based distance functions

Θεωρούν δύο string ως δύο σύνολα λέξεων, έστω S και T. Μέθοδοι αυτής της κατηγορίας αποτελούν οι:

1.3.2.1 $Jaccard\_similarity = \frac{|S \cap T|}{|S \cup T|}$

1.3.2.2 $TFIDF(S,T) = \sum_{w \in S \cap T} V(w,S)V(w,T)$ η οποία χρησιμοποιείται ευρέως σε εφαρμογές ανάκτησης πληροφορίας ,όπου $TF_{w,S}$ η συχνότητα της w λέξης στο S, IDF_w το αντίστροφο κλάσμα των ονομάτων του συνόλου που περιέχουν την w, $V'(w,S) = \log(TF_{w,S}+1)\log(IDF_w)$ και $V(w,S) = V'(w,S) \Big/ \sqrt{\sum_{w'} V'(w,S)^2}$

1.3.2.3 $\text{Jensen - Shannon}(S, T) = \frac{1}{2}(KL(P_S \parallel Q) + KL(P_T \parallel Q))$,όπου τα σύνολα συμβόλων S και T αποτελούν δείγματα από άγνωστες κατανομές συμβόλων P_S κα P_T, $KL(P \parallel Q)$ η απόκλιση Kullback-Lieber και $Q(w) = \frac{1}{2}(P_S(w) + P_T(w))$

1.3.2.4 Fellegi and Sunter [34]

1.3.3 Hybrid distance functions

Η μέθοδος Monge and Elkan [35],[36], αντιμετωπίζει με αναδρομικό τρόπο την σύγκριση δύο μεγάλων string s και t, σπάζοντάς τα αρχικά σε μικρότερα string $s = a_1...a_k$ και $t = b_1...b_L$. Η ομοιότητά τους ορίζεται ως $sim(s,t) = \frac{1}{K}\sum_{i=1}^{K} \max_{j=1}^{L}(sim'(A_i, B_j))$,όπου sim' βοηθητική distance function.

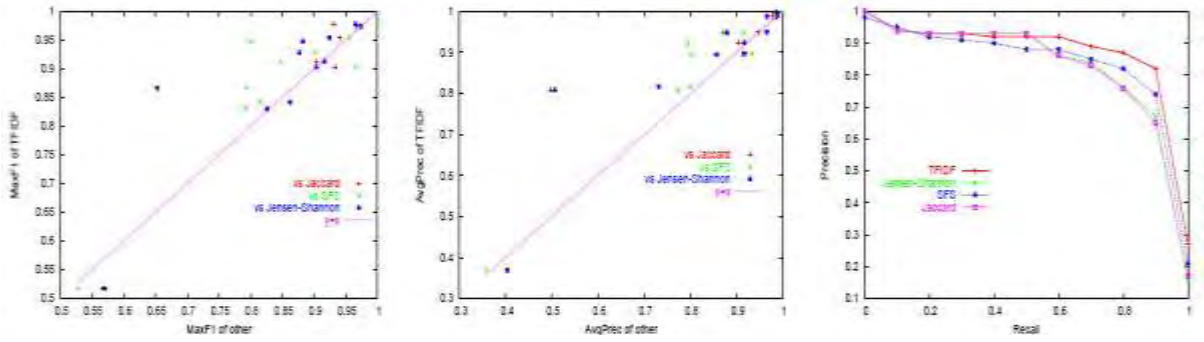
Άλλη μία αξιόλογη μέθοδος αυτής της κατηγορίας είναι η SoftTFIDF στην οποία παρόμοιοι χαρακτήρες θεωρείται πως ανήκουν στο σύνολο $S \cap T$ ορίζεται ως

$SoftTFIDF(S,T) = \sum_{w \in CLOSE(\theta,S,T)} V(w,S) * V(w,T) * D(w,T)$, όπου sim' βοηθητική distance function, $CLOSE(\theta,S,T)$ πλήθος w λέξεων, $w \in S$ ώστε $dist'(w,u) > \theta, u \in T$ και για κάθε $w \in CLOSE(\theta,S,T)$ θεωρούμε πως $D(w,T) = \max_{u \in T} dist(w,u)$.

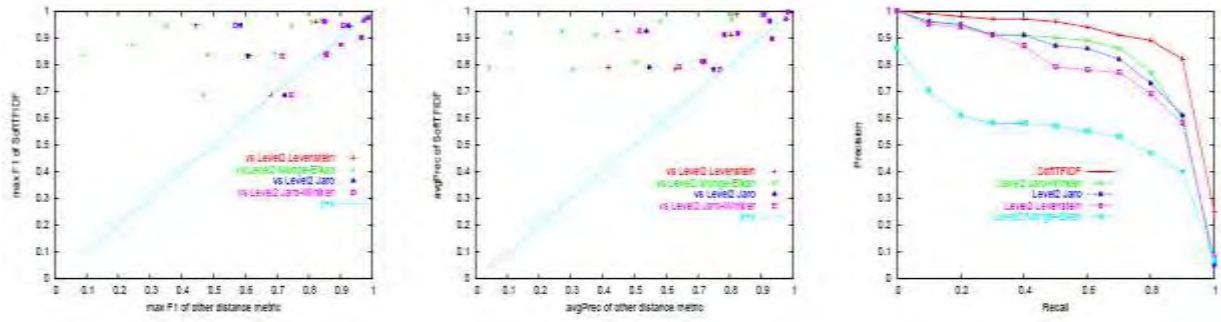
Τέλος αναφέρονται οι Blocking methods οι οποίες κατά την αντιστοίχιση μεγάλων λιστών δεδομένων, όπου δεν είναι υπολογιστικά πρακτικό να μετρηθούν οι αποστάσεις μεταξύ όλων των ζευγών string, λαμβάνονται υπόψη μεταβλητές που είναι συνήθως ίδιες για δεδομένα παρόμοιου τύπου.

1.3.4 Πειραματικά Αποτελέσματα

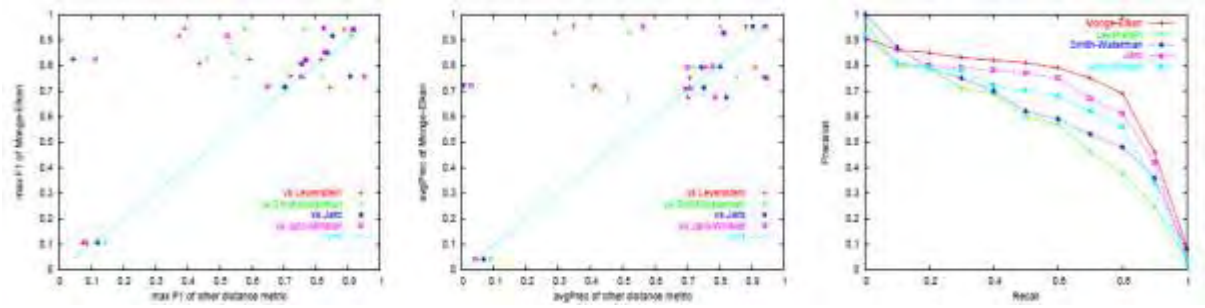
Τα datasets στα οποία εφαρμόζονται οι προηγούμενοι μέθοδοι αντιμετωπίζουν δύο κατηγορίες προβλημάτων. Η πρώτη κατηγορία αναφέρεται σε προβλήματα αντιστοίχισης ενώ η δεύτερη σε προβλήματα ομαδοποίησης. Όσο αφορά την αντιστοίχιση, από τις token based distance μεθόδους ο TFIDF έχει την καλύτερη επίδοση (εικόνα 4), από τις hybrid ο SoftTFIDF (εικόνα 5) ενώ για τις edit distance την καλύτερη μέση επίδοση την έχει ο Monge-Elkan, αλλά οι Jaro με τις παραλλαγές της, είναι πολύ κοντά σε αυτά τα αποτελέσματα (εικόνα 6) ενώ παρουσιάζουν χρόνο εκτέλεσης ίσο με το 1/10 της πρώτης.



Εικόνα 4: Σύγκριση των token based distance μεθόδων. Αριστερά η μέγιστη τιμή F1 κάθε μεθόδου σε σχέση με αυτή της TFIDF όπου τα σημεία πάνω από την $y = x$ να υποδηλώνουν καλύτερη λειτουργία της TFIDF. Αντίστοιχα στο μεσαίο σχήμα για την μέση ακρίβεια ενώ δεξιά εμφανίζεται η σχέση precision – recall.

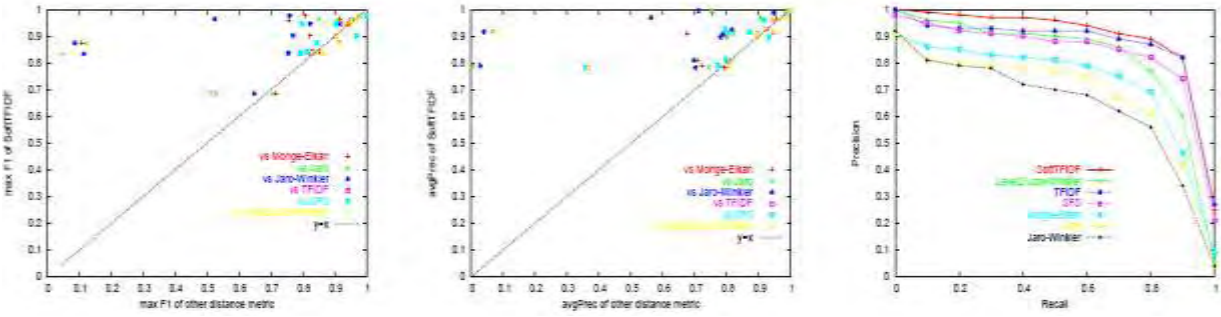


Εικόνα 5: Σχετική απόδοση των hybrid μεθόδων.



Εικόνα 6: Σχετική απόδοση των edit distance μεθόδων.

Συγκρίνοντας τις αποδοτικότερες μεθόδους κάθε κατηγορίας, ο SoftTFIDF αναδεικνύεται αποτελεσματικότερος (εικόνα 7).



Εικόνα 7: Σύγκριση αποδοτικότερων μεθόδων

Στο πρόβλημα ομαδοποίησης σύμφωνα με τον πίνακα 1, ο SoftTFIDF αναδεικνύεται και πάλι αποτελεσματικότερος.

Method	MaxF1	AvgPrec	MaxF1	AvgPrec
SoftTFIDF	0.89	0.91	0.85	0.914
TFIDF	0.79	0.84	0.84	0.907
SFS	0.71	0.75	0.82	0.864
Level2 J-W	0.73	0.69	0.76	0.804

Πίνακας 1: Αποτελέσματα μεθόδων κατά την ομαδοποίηση

Αν συνδυάσουμε στοιχεία από όλες τις διαθέσιμες μεθόδους μπορεί να παραχθούν νέες με ακόμη καλύτερη συμπεριφορά. Τα αποτελέσματα των πειραμάτων διεξήχθησαν πάνω σε συγκεκριμένους τύπους δεδομένων και ίσως να υπάρξουν διαφορετικά αποτελέσματα σε άλλες περιπτώσεις.

1.4 Conclusions

Σε αυτό το κεφάλαιο μελετήσαμε με ποιον τρόπο μπορούμε να ανακτήσουμε αυτόματα πληροφορίες δομημένης μορφής από ιστοσελίδες με ιδιαίτερα υψηλό ποσοστό επιτυχίας. Κατά την μελέτη των αλγορίθμων που προσεγγίζουν την λύση του προβλήματος της εξόρυξης, παρατηρήθηκε η αναγκαιότητα της χρήσης μεθόδων που ελέγχουν την ομοιότητα μεταξύ δύο string. Επειδή η ποιότητα αυτού του ελέγχου έχει άμεση σχέση με την ποιότητα της εξόρυξης μελετήσαμε ορισμένες τέτοιες μεθόδους ώστε να καταλάβουμε την λογική με την οποία προσεγγίζει το πρόβλημα η καθεμία αλλά και να αποφασίσουμε για το ποια επιστρέφει καλύτερα αποτελέσματα.

2. Citation matching/resolution

2.1 Introduction

Το γεγονός ότι ο λόγος μπορεί να χρησιμοποιηθεί με διαφορετικό τρόπο από τον κάθε άνθρωπο μας βοηθάει να συμπεράνουμε πως ένα αντικείμενο μπορεί να έχει πολλούς διαφορετικούς τρόπους περιγραφής. Στην περίπτωση όμως των παραπομπών ενός paper, από τα οποία ο αναγνώστης θέλει με σιγουριά να γνωρίζει σε ποια paper αναφέρονται, κάτι τέτοιο αποτελεί πρόβλημα. Αυτό γιατί στον κόσμο των υπολογιστών δεν είναι δυνατό να υπάρξει τέλειος αλγόριθμος που να προσπελαύνει ένα ελεύθερο κείμενο και να αποφαίνεται με απόλυτη επιτυχία πως μία παραπομπή αντιστοιχίζει σε ένα συγκεκριμένο paper. Ρεαλιστικά όμως μπορούμε να είμαστε ικανοποιημένοι και με προσεγγιστικές λύσεις οι οποίες θα εκτελούν την αντιστοίχιση με το ελάχιστο δυνατό σφάλμα. Παρακάτω προτείνονται μέθοδοι που προσπαθούν να επιτύχουν τέτοιες ελαχιστοποιήσεις τις οποίες και θα μελετήσουμε.

2.2 Autonomous Citation Matching

Το παρόν paper μας προτείνει την ανάπτυξη ενός αυτόνομου συστήματος αρχειοθέτησης παραπομπών το οποίο σε αντίθεση με τα παραδοσιακά συστήματα δεν απαιτεί ανθρώπινη προσπάθεια πράγμα που το καθιστά λειτουργικότερο και ευκολότερα προσβάσιμο στο κοινό. Γενικότερα οι παραπομπές ενός επιστημονικού άρθρου μπορούν να αναπαρασταθούν με πολλές διαφορετικές μορφές για παράδειγμα:

-Rosenblatt F. (1961). Principle o Neurodynamics: Perceptrons and the Theory of Brain Mechanisms. Spartan Books, Washington, D.C.
-Rosenblatt, F. (1962). Principle o Neurodynamics. Washington, DC: Spartan.
-F. Rosenblatt. Principle o Neurodynamics. Spartan Books, 1962.

Μπορούμε εύκολα να αναλογιστούμε την υψηλή πολυπλοκότητα του προβλήματος μιας και δεν υπάρχει κάποιος προφανής τρόπος να ορίσουμε τα μεταδεδομένα μιας παραπομπής. Η πιθανή ύπαρξη ορθογραφικών λαθών προφανώς κάνει το πρόβλημα ακόμα πιο δύσκολο. Στην προσπάθειά να αποσαφηνιστούν ποιες παραπομπές αντιστοιχούν στα ανάλογα paper, μπορούν να χρησιμοποιηθούν τέσσερα είδη μεθόδων που εκτελούν τις παρακάτω λειτουργίες:

a) Έλεγχος ομοιότητας δύο string με βάση τη μεταξύ τους απόσταση.
b) Έλεγχος συχνότητας εμφάνισης μιας λέξης.
c) Αξιοποίηση πληροφοριών που περιέχονται σε μια παραπομπή όπως ο τίτλος και οι συγγραφείς του paper, το έτος δημοσίευσης κλπ .
d) Χρήση μοντέλων πιθανοτήτων με σκοπό την εύρεση των υποπεδίων μιας παραπομπής μέσα από συγκεκριμένες λέξεις ή/και τον τρόπο δόμησής της.

Καλό θα ήταν οι παραπομπές αρχικά να υπόκεινται κάποια στάδια επεξεργασίας ώστε να μετατραπούν σε κάποια κανονικοποιημένη μορφή. Τέτοια στάδια είναι η μετατροπή όλων των κεφαλαίων γραμμάτων σε μικρά, η μετατροπή κάθε παύλας σε κενό, η αφαίρεση άχρηστων πληροφοριών όπως περιττές λέξεις και ετικέτες παραπομπών και η ερμηνεία των γνωστότερων συντομογραφιών.

Προτείνονται στο paper οι εξής αλγόριθμοι:

2.3.1 ***Simple Baseline Algorithm*** **-** Ομαδοποιεί τις παραπομπές που θεωρεί πως αναφέρονται στο ίδιο paper με τη λογική πως αυτές που μοιράζονται ένα ποσοστό κοινών λέξεων μεγαλύτερο από κάποια ορισμένη τιμή, τότε θα ταυτίζονται.

Αλγόριθμος 3: Simple Baseline Algorithm

For each citation a
If this citation has not been grouped yet
Create a new group with this citation
For each remaining ungrouped citation b
Add citation b to the current group if the number of the matching words..
Is greater than x% of the length of the shorter citation
End for
End if
End for

2.3.2 ***Word Matching Algorithm*** – Παρόμοια λογική με τον προηγούμενο αλγόριθμο (i) με την διαφορά πως πρώτα ταξινομούμε κάθε ομάδα παραπομπών για να χρησιμοποιούμε την μεγαλύτερη ως αναγνωριστικό της, αφού λογικά θα περιέχει και τις περισσότερες πληροφορίες για το paper που αναφέρεται. Εντάσσουμε στην συνέχεια την παραπομπή στην ομάδα με το μεγαλύτερο ποσοστό αντιστοιχίας αν αυτό βέβαια ξεπερνά την ελάχιστη τιμή που ορίσαμε, αλλιώς δημιουργούμε νέα ομάδα. Ακολουθεί ταξινόμηση και έλεγχος μεταξύ των ομάδων, οι οποίες πιθανώς λανθασμένα να διαχωρίστηκαν στο παρελθόν. Η παρακάτω υλοποίηση είναι βασισμένη σε hash table λέξεων όπου κάθε καταχώρηση περιέχει μία λίστα με τις ομάδες που περιέχουν την λέξη αυτή.

Αλγόριθμος 4: Word Matching Algorithm
Sort the citations by length, from the longest to the shortest citation For each citation c For each word in the citation Find the array of groups containing this word from the words hash table End for Find the most common group g in the arrays of groups for each word If the ratio of the number of non-matching words to the matching words in c..

```
            is less than a threshold then add c to the group g
        Else create a new group for this citation by adding the new group number to the…
            arrays of groups for each word, which are contained in the words hash table
        End if
End for
Order the groups such that each successive group has the most number of words in..
  common with the previous group out of all following groups
```

2.3.3 ***Word and Phrase Matching Algorithm*** – Σχεδόν πανομοιότυπη λογική και υλοποίηση με τον αλγόριθμο (ii) με τη διαφορά ότι λαμβάνει επιπλέον υπόψη φράσεις δύο λέξεων που περιέχονται σε τμήματα διαχωρισμένα από ‘,’ και ‘.’ σύμβολα, που περιέχουν τρείς ή περισσότερες λέξεις.

Αλγόριθμος 5: Word and Phrase Matching Algorithm

```
Sort the citations by length, from the longest to the shortest citation
For each citation c
        Find the group g with the highest number of matching words
        Let a = ratio of the number of non-matching words to the number of matching
words
        Let b = ratio of the number of non-matching phrases to the number of matching…
            phrases where a phrase is every set of two successive in every section of the…
            citation containing 3 or more words (section delimited by ., or ..)
        If (a < threshold1) or (a < threshold2 and b < threshold3)
        Then add c to the group g
        Else create a new group for this citation
End for
Order the groups such that each successive group has the most number of words in..
  common with the previous group out of all following groups
```

2.3.4 ***LikeIt Edit Distance Algorithm*** – Ελέγχει κατά πόσο δύο string είναι όμοια. Όσο λιγότερες προσθαφαιρέσεις και μετακινήσεις λέξεων χρειάζεται ένα string για να γίνει πανομοιότυπο με ένα άλλο, τόσο περισσότερο μοιάζουνε.

Αλγόριθμος 6: Algorithm based on LikeIt

Sort the citations by length, from the longest to the shortest citation
For each citation c
Find the group g with the highest value d = LikeIt(c , c)/LikeIt(c , g)
If d is greater than a threshold then add c to the group g
Else create a new group for this citation
End if
End for
Order the groups such that each successive group has the most number of words in..
common with the previous group out of all following groups

2.3.5 ***Subfield Algorithm*** – Αλγόριθμος που στοχεύει στην ανάκτηση και εκμετάλλευση πληροφοριών από μία παραπομπή όπως ο τίτλος, οι συγγραφείς κτλ. Βασίζεται σε παρατηρήσεις που γίνονται στην δομή των παραπομπών μιας και ο διαχωρισμός και η εύρεση συγκεκριμένων πληροφοριών από ένα σχετικά ελεύθερο κείμενο αποτελεί μεγάλη πρόκληση.

2.3.6 <u>Πειραματικά αποτελέσματα</u>

Στα πειράματα που διεξήχθησαν λήφθηκαν υπόψη 1.947 papers εκμάθησης μηχανής κατεβασμένα από το διαδίκτυο, τα οποία αποτελούνταν από 39.166 παραπομπές. Από αυτές δημιουργήθηκαν τέσσερις νέες υποκατηγορίες παραπομπών έτσι ώστε κάθε ομάδα να περιέχει τις λέξεις “reinforcement”, “constraint”, “face” και “reasoning” με την καθεμία να περιέχει 406, 295, 349 και 514 παραπομπές αντίστοιχα. Αρχικά η ταξινόμηση έγινε χειροκίνητα προκειμένου να αποκτηθούν τα σωστά αποτελέσματα για τον σωστό τρόπο ταξινόμησης των παραπομπών, έτσι ώστε να μπορέσουμε να τα συγκρίνουμε με αυτά των αλγορίθμων για να εξάγουμε τα τελικά μας συμπεράσματα. Οι πίνακες 2 και 3 περιγράφουν τα ποσοστά των εσφαλμένων αποτελεσμάτων κατά την αυτόματη ομαδοποίηση.

	Constraint	Face	Reasoning	Average
Number of citations	295	349	514	
Baseline Simple	12%	6%	13%	10.3%
Word Matching	9%	4%	8%	7%
Word and Phrase Matching	**6%**	**3%**	**7%**	**5.3%**
LikeIt	13%	13%	17%	14.3%
Subfield	12%	9%	16%	12.3%

Πίνακας 2: Τα αποτελέσματα των αυτόματων αλγορίθμων αντιστοίχισης σε κανονικοποιημένα δεδομένα

	Constraint	Face	Reasoning	Average
Number of citations	295	349	514	
Baseline Simple	16%	12%	19%	15.7%
Word Matching	16%	11%	20%	15.7%
Word and Phrase Matching	14%	9%	17%	13.3%
LikeIt	9%	10%	14%	11%
Subfield	12%	9%	16%	12.3%

Πίνακας 3: Τα αποτελέσματα των αυτόματων αλγορίθμων αντιστοίχισης σε μη κανονικοποιημένα δεδομένα

Με κανονικοποίηση ο Word and Phrase Matching αλγόριθμος παρουσιάζει το χαμηλότερο ποσοστό μέσου σφάλματος, ενώ χωρίς κανονικοποίηση ο LikeIt αλγόριθμος θεωρείται αποτελεσματικότερος. Συντριπτικά ταχύτερος όλων αποτελεί ο Subfield αλγόριθμος.

2.3 Identity Uncertainty and Citation Matching

Η αβεβαιότητα ταυτότητας προκύπτει από το γεγονός πως κάθε αντικείμενο μπορεί να περιγραφεί με περισσότερο από ένα τρόπους και πάντοτε με κάποια πιθανότητα λάθους. Το φαινόμενο αυτό παρατηρείται και στο πρόβλημα αντιστοίχισης παραπομπών με τα αντίστοιχα επιστημονικά άρθρα. Η παρακάτω προσέγγιση χειρίζεται το πρόβλημα αυτό επεκτείνοντας τυποποιημένα μοντέλα με σκοπό να ενσωματώσει πιθανότητες κατά την αντιστοίχιση των όρων της γλώσσας με τα αντικείμενα στα οποία αναφέρονται. Τα συμπεράσματα λαμβάνονται βάση της μεθόδου Markov Chain Monte Carlo, επεκταμένη με τέτοιο τρόπο ώστε να παράγει αποτελεσματικές προτάσεις όταν ένας τομέας περιέχει πολλά αντικείμενα.

Εισάγουμε την έννοια του συσχετιστικού μοντέλου πιθανοτήτων (RPM)[26],[27], το οποίο μας επιτρέπει να προσδιορίσουμε μοντέλα πιθανοτήτων γύρω από πιθανούς κόσμους ορισμένους από κλάσεις, αντικείμενα, ιδιότητες και σχέσεις.

2.3.1 Γενικά ένα RPM αποτελείται από:

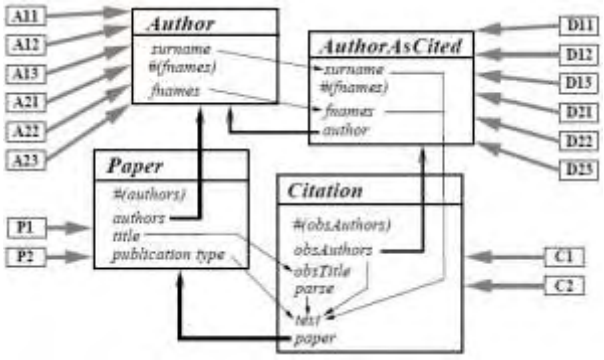
a) Μία συλλογή C κλάσεων που συσχετίζονται με υποκλάσεις και υπερκλάσεις και δηλώνει πλήθος αντικειμένων.
b) Μία συλλογή I ονομασμένων στιγμιότυπων κλάσεων που υποδηλώνει αντικείμενα.

c) Μία συλλογή A σύνθετων χαρακτηριστικών που δηλώνουν σχέσεις μεταξύ συναρτήσεων με την καθεμία να περιέχει ένα τύπο τομέα $Dom[A] \in C$ και εμβέλεια $Range[A] \in C$.

d) Μία συλλογή B απλών χαρακτηριστικών που δηλώνουν συναρτήσεις με την καθεμία να περιέχει ένα τύπο τομέα $Dom[B] \in C$ και εμβέλεια $Range[B] \in C$.

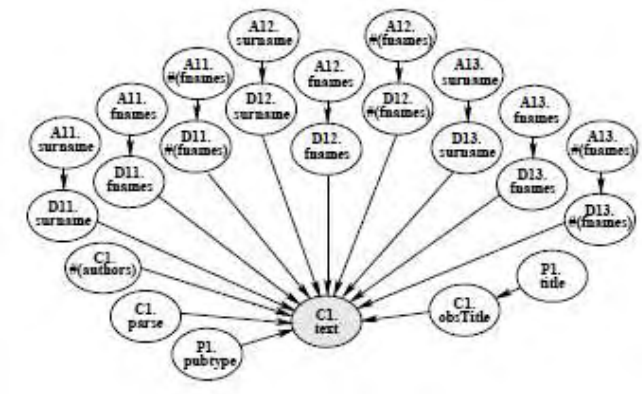
e) Μία συλλογή από υπό όρους μοντέλα πιθανοτήτων $P(B \mid P_a[B])$ που αφορούν τα απλά χαρακτηριστικά με $P_a[B]$ να αποτελεί το σύνολο των γονέων του B, καθένα από τα οποία είναι μία μη κενή αλυσίδα χαρακτηριστικών B'

f) Μία συλλογή από δηλώσεις στιγμιότυπων που ορίζουν την τιμή ενός σύνθετου χαρακτηριστικού στο στιγμιότυπο της κατάλληλης κλάσης.



Εικόνα 8: RPM παραπομπής, όπου τα μεγάλα ορθογώνια υποδηλώνουν κλάσεις, τα μαύρα βέλη εμβέλειες των σύνθετων χαρακτηριστικών, τα γκρι βέλη τις πιθανολογικές εξαρτήσεις των βασικών χαρακτηριστικών ενώ τα μικρά ορθογώνια στιγμιότυπα κλάσεων

Σε ένα RPM θεωρούμε πως κάθε όνομα είναι μοναδικό με αποτέλεσμα κάθε paper να θεωρείται διαφορετικό. Μπορούμε να εκφράσουμε ένα RPM ως ένα ισοδύναμο δίκτυο Bayes, όπως αυτό της εικόνας 9.κάθε κόμβος αντιστοιχεί σε ένα βασικό χαρακτηριστικό



Εικόνα 9: Μορφή δικτύου Bayes κάθε κόμβος αντιστοιχεί σε ένα βασικό χαρακτηριστικό

Όταν έχουμε δύο παραπομπές C_1 και C_2 υπάρχει ανακρίβεια για το αν αντιστοιχούν σε δύο ίδια paper ή όχι. Αν είναι ίδια τότε θα μοιράζονται ένα κοινό σύνολο βασικών χαρακτηριστικών ενώ στην αντίθετη περίπτωση θα υπάρχουν δύο διαφορετικά σύνολα χαρακτηριστικών. Έτσι οι πιθανοί κόσμοι από το μοντέλο πιθανοτήτων μπορεί να διαφέρουν στον αριθμό των τυχαίων μεταβλητών. Κάθε κόσμος επεκτείνεται έτσι ώστε να περιλαμβάνει επιπλέον το πλήθος n των αντικειμένων και μία ταυτότητα ομαδοποίησης i η οποία θα αντιστοιχίζει τους όρους μιας γλώσσας στα αντικείμενα του κόσμου. Για παράδειγμα αν δύο paper αναφέρονται στο ίδιο αντικείμενο τότε το i θα είναι $\{P_1, P_2\}$, ενώ στην αντίθετη περίπτωση θα ισούται με $\{\{P_1\},\{P_2\}\}$.

Το μοντέλο πιθανοτήτων που ορίζεται για τον χώρο των επεκταμένων πιθανών κόσμων λαμβάνει υπόψη την πιθανότητα $P(n)$ του προηγούμενου κόσμου και την υπό συνθήκη κατανομή $P(i \mid n)$ για τον καθένα, τις οποίες μπορούμε να απλοποιήσουμε παραγοντοποιώντας τες ανά κλάση. Κατά την παραγοντοποίηση διακρίνουμε δύο περιπτώσεις.

a) Ορίζουμε πιθανότητα ίση με την μονάδα σε κλάσεις όπως οι παραπομπές, τα αντικείμενα των οποίων θεωρούνται μοναδικά ($P(i_{Citation}) = 1$)
b) Όταν τα στοιχεία των κλάσεων υπόκεινται από αβεβαιότητα ταυτότητας, ορίζεται αρχικά η $P(n)$ με τη χρήση μιας υψηλής διακύμανσης κατανομή και στην συνέχεια, αφού υποθέσουμε ότι κάθε αντικείμενο πριν από τη χορήγηση οποιαδήποτε πληροφορίας είναι εξίσου πιθανό να αναφερθεί, κατασκευάζεται η $P(i_C)$. Βάσει αυτής της υπόθεσης η πιθανότητα της $i_{C,k,m}$,η οποία αντιστοιχίζει k στιγμιότυπα σε m αντικείμενα με την κλάση C να περιέχει n αντικείμενα, δίνεται από την σχέση $P(i_{C,k,m}) = \frac{n!}{(n-m)!}\frac{1}{n^k}$. Αν $n > m$, ο κόσμος περιέχει άσχετα αντικείμενα τα οποία και θα πρέπει να αγνοηθούν.

Το μοντέλο αυτό υποθέτει ανεξαρτησία μεταξύ των κλάσεων και των χαρακτηριστικών η οποία αν δεν θεωρηθεί δεδομένη θα δημιουργήσει αυτόματα εξαρτήσεις οι *οποίες* μπορούν να καθοριστούν από ένα δίκτυο Bayes. Για παράδειγμα το πλήθος των paper εξαρτάται από το πλήθος των συγγραφέων.

Εισάγοντας την έννοια της αβεβαιότητας το απλό δίκτυο Bayes μετατρέπεται σε μία συλλογή δικτύων, μία για κάθε δυνατή ομάδα i, με αποτέλεσμα η ακριβής εξαγωγή συμπερασμάτων να είναι αδύνατη. Σε αυτή την περίπτωση χρησιμοποιείται μία μέθοδος προσέγγισης βασισμένη στην *Markov Chain Monte Carlo* μέθοδο (MCMC)[29]. Η MCMC μέθοδος προσεγγίζει μία προσδοκώμενη τιμή μιας κατανομής $\pi(x)$, με τον χώρο καταστάσεων του x να είναι πολύ μεγάλος. Το άθροισμα των τιμών του x αντικαθίσταται με το άθροισμα δειγμάτων από την $\pi(x)$ τα οποία παράγονται με την βοήθεια των αλυσίδων Markov. Υπάρχουν διάφοροι τρόποι για την δημιουργία κατάλληλης αλυσίδας Markov. Σε έναν από

αυτούς οι μεταβάσεις στην αλυσίδα κατασκευάζεται σε δύο στάδια. Πρώτον, μια υποψήφιος επόμενη κατάσταση x' παράγεται από την τρέχουσα κατάσταση x, χρησιμοποιώντας την προτεινόμενη τιμή $q(x'|x)$, με την πιθανότητα μετακίνησης στο x' να είναι ίση με

$$a(x'|x) = \min\left(1, \frac{\pi(x')q(x|x')}{\pi(x)q(x'|x)}\right).$$

Στον MCMC αλγόριθμο που παρουσιάζεται ο χώρος καταστάσεων του είναι ο χώρος όλων των πιθανών κόσμων, με κάθε κόσμο να περιέχει μία ομάδα ταυτοτήτων i, ένα σύνολο από n πληθικότητες και τις τιμές από όλα τα χαρακτηριστικά όλων των αντικειμένων. Αρχικά προτείνεται μια αλλαγή στον i και στη συνέχεια υπολογίζει τις αλλαγές για όλα τα κρυμμένα χαρακτηριστικά όλων των αντικειμένων που επηρεάζονται από την αλλαγή αυτή.
Ο αλγόριθμος αυτός λειτουργεί με τον εξής τρόπο:

Αλγόριθμος 7: MCMC
Διάλεξε δύο ομάδες $a_1, a_2 \in i_C$ Δημιούργησε δύο άδειες ομάδες b_1 και b_2 Τοποθέτησε κάθε στιγμιότυπο $i \in a_1 \cup a_2$ στην b_1 ή στην b_2 Πρότεινε $i'_C = i_C - \{a_1, a_2\} \cup \{b_1, b_2\}$

Μετά από μία πρόταση i'_C γίνεται προσπάθεια ανάκτησης των πραγματικών τιμών κάθε χαρακτηριστικού σε περίπτωση εσφαλμένης παρατήρησης.
Ο προηγούμενος αλγόριθμος αντιμετωπίζει προβλήματα κατά την κλιμάκωση του προβλήματος, μιας και όσο αυξάνουμε τον αριθμό των paper τόσο είναι πιθανότερο οι προτάσεις να απορρίπτονται, με αποτέλεσμα η αλυσίδα Markov να συντίθεται πολύ αργά.

2.3.2 Πειραματικά αποτελέσματα

Αξιολογούμε τον MCMC αλγόριθμο πάνω στο dataset που χρησιμοποιήθηκε στο [2] και τον συγκρίνουμε με τον phrase matching αλγόριθμο που περιγράφεται εκεί. Παρατηρούμε στον πίνακα 4 την υπεροχή που παρουσιάζει στην αντιστοίχιση ο MCMC αλγόριθμος έναντι του δεύτερου.

	Face 349 citations, 242 papers	Reinforcement 406 citations, 148 papers	Reasoning 514 citations, 296 papers	Constraint 295 citations, 199 papers
Phrase matching	94%	79%	86%	89%
RPM + MCMC	97%	94%	96%	93%

Πίνακας 4: Ποσοστά επιτυχίας αντιστοίχισης του MCMC και του phrase matching αλγορίθμου

2.4 Constraint-Based Entity Matching

Κάθε εφαρμογή στον πραγματικό κόσμο, περιέχει πολλούς σημασιολογικούς περιορισμούς ακεραιότητας. Κατά την αντιστοίχισης οντοτήτων, η εισαγωγή λογικών περιορισμών οδηγεί σε μια έξυπνη και συνεπώς αποτελεσματικότερη προσέγγιση του προβλήματος.

Για παράδειγμα το γεγονός ότι δεν υπάρχει συγγραφέας που να έχει δημοσιεύσει περισσότερο από πέντε AAAI papers σε ένα χρόνο μας οδηγεί στο συμπέρασμα πως αν δούμε το ίδιο όνομα να αναφέρεται σε έξι ή περισσότερα papers του ίδιου έτους, τότε δεν θα πρόκειται για το ίδιο φυσικό πρόσωπο σε όλες τις περιπτώσεις.

Οι περιορισμοί μπορούν να οριστούν από τα δεδομένα, ή να καθορίζονται από έναν εμπειρογνώμονα/χρήστη και μπορούν να χαρακτηριστούν ως soft ή hard. Soft χαρακτηρίζονται οι περιορισμοί που η ικανοποίησή τους δεν μπορεί να μας εγγυηθεί την εξαγωγή απόλυτων συμπερασμάτων και μπορούν να πάρουν τιμές από [0,..,1], ενώ hard περιορισμούς χαρακτηρίζουμε αυτούς που η ικανοποίηση τους ή μη έχει πιο απόλυτο χαρακτήρα, και παίρνουν τιμές 1 και 0 αντίστοιχα. Το προαναφερόμενο παράδειγμα αποτελεί παράδειγμα hard περιορισμού, ενώ ως soft μπορούμε να χαρακτηρίσουμε τον περιορισμό, αν δύο παραπομπές έχουν έναν αριθμό από κοινούς συγγραφείς τότε αυτά είναι ίδια με πιθανότητα έστω 90%. Η τιμή που περιγράφει έναν soft περιορισμό δεν είναι απόλυτη και εξαρτάται από την κρίση και αντίληψη αυτού που την ορίζει.

Η CME μέθοδος που μας προτείνεται για την αντιμετώπιση του προβλήματος της αντιστοίχισης οντοτήτων, βασίζεται σε μια αρχιτεκτονική δύο επιπέδων όπου το πρώτο εκμεταλλεύεται περιορισμούς σε ομαδικό επίπεδο και ομαδοποιεί τις παραπομπές έτσι ώστε όλες οι παρόμοιες να ανήκουν στην ίδια ομάδα. Πιο συγκεκριμένα εφαρμόζεται επαναληπτικά ένας συνδυασμός του EM αλγορίθμου, που υπολογίζει τις παραμέτρους ενός παραγωγικού μοντέλου και εκτελεί αντιστοιχίσεις και του αλγορίθμου χαλάρωσης ετικετών (relaxation labeling), που εκμεταλλεύεται τους περιορισμούς προκειμένου να βελτιώσει την ακρίβεια των υπολογισμών. Το παραγωγικό μοντέλο προκειμένου να δημιουργήσει μία ομάδα δεδομένων D, παράγει τα έγγραφα $d_1, d_2, .., d_n$. Για να παραχθεί το d_1 επιλέγεται ένας τυχαίος αριθμός num_e οντοτήτων E_1 από όλες τις πιθανές οντότητες E σύμφωνα με μία πιθανότητα $P(E_1)$. Για κάθε οντότητα $e \in E_1$ παράγονται num_m παραπομπές m, οι οποίες αποθηκεύονται τυχαία στο έγγραφο d_1, με την καθεμία να παράγεται ανεξάρτητα από το e σύμφωνα με μία μετασχηματισμένη πιθανότητα $P(m \mid e)$. Όταν ολοκληρωθεί η δημιουργία

της ομάδας δεδομένων $D = d_1, d_2, .., d_n$ ελέγχεται αν ικανοποιεί ένα πλήθος περιορισμών C που χαρακτηρίζονται από ανάλογες πιθανότητες. Αν ναι, τότε το σύνολο δεδομένων διατηρείται, αλλιώς απορρίπτεται και η διαδικασία επαναλαμβάνεται. Ορίζουμε ως σύνολο παραμέτρων θ του παραγωγικού μοντέλου τις εξής:

- Ένα σετ οντοτήτων E_D και μία κατανομή $P(E_D)$ που να την περιγράφει
- Κατανομές πιθανοτήτων σχετικά με τους num_e και num_m
- Πιθανότητες παραγωγής παραπομπών $P(m \mid e)$
- Πιθανότητες $p_1, ..., p_l$ για τους περιορισμούς $c_1, ..., c_l$

Αν F αποτελεί μία ανάθεση από παραπομπές της D σε οντότητες, τότε εμείς ενδιαφερόμαστε για τον υπολογισμό των βέλτιστων F^* και θ^* ώστε $P(D, F^* \mid \theta^*)$ να έχει μέγιστη τιμή. Βάση της F^* μπορούμε να αποφασίσουμε για το ποιες παραπομπές αναφέρονται στις ίδιες οντότητες. Όμως εκθετικός αριθμός των πιθανών F και θ καθιστούν την εύρεση των F^* και θ^* ως μία ασύμφορη και μη-ρεαλιστική διαδικασία. Παρακάτω περιγράφεται μία παραλλαγή του ΕΜ αλγορίθμου που υπολογίζει αυτές τις δύο τιμές.

Αλγόριθμος 8: Παραλλαγή ΕΜ αλγορίθμου
1) (Initialization): Έστω $t \leftarrow 0$. Βρες μια αρχική ανάθεση F_0, η οποία αναθέτει κάθε αναφορά στο D σε μία οντότητα του πραγματικού κόσμου.
2) (Maximization): Υπολόγισε τις παραμέτρους $\theta_{t+1} \leftarrow \arg\max_\theta P(D, F_t \mid \vartheta)$.
3) (Expectation): Υπολόγισε τις αναθέσεις των παραπομπών $F_{t+1} \leftarrow \arg\max_F P(D, F \mid \vartheta_{t+1})$
4) (Convergence): Για μία τιμή ε, If $[P(D, F_{t+1} \mid \vartheta_{t+1}) - P(D, F_t \mid \vartheta_t)] \le \varepsilon \quad \{return(F_{t+1})\}$ else $t \leftarrow t+1$ and go to 2)

Στο επίπεδο που περιγράφηκε, δύο παραπομπές είναι παρόμοιες όταν αναφέρονται στην ίδια οντότητα. Αν ένα σετ παραπομπών αναφέρεται στην ίδια οντότητα, υπονοείται πως υπάρχει ισοδυναμία μεταξύ κάθε δυνατού ζεύγους παραπομπών. Αν υπάρχουν περιορισμοί που εφαρμόζονται σε μικρό υποσύνολο ζευγών από παραπομπές, τότε σε γενικές γραμμές υπάρχει κίνδυνος να επηρεάσουν αρνητικά την συνολική ακρίβεια του αλγορίθμου. Για τον λόγο αυτό προστίθεται ένα δεύτερο επίπεδο που αναλαμβάνει να αξιοποιήσει πρόσθετους περιορισμούς

σε επίπεδο επιμέρους ζευγών. Θα εξετάζει δηλαδή εννοιολογικά όλα τα ισοδύναμα ζεύγη παραπομπών που επέστρεψε το πρώτο στρώμα εφαρμόζοντας αυστηρούς περιορισμούς και μετά θα επιστρέφει τα τελικά αποτελέσματα στον χρήστη. Για παράδειγμα ένα πρόσωπο ηλικίας δύο ετών δεν θα μπορεί να ταιριάζει με ένα πρόσωπο με ετήσιες απολαβές 200K. Ο χρήστης στο τέλος αφού μελετήσει τα αποτελέσματα της αυτόματης εκτέλεσης του αλγορίθμου, μπορεί να εισάγει δικούς του περιορισμούς και να ξανατρέξει την μέθοδο από τον relaxation labeling αλγόριθμο για να δει πως τελικά αυτοί ανταποκρίνονται.

- Πειραματικά αποτελέσματα

Στα πειράματα έγινε χρήση δύο διαφορετικών data sets. Συγκεκριμένα το Researchers περιλαμβάνει 4.991 παραπομπές, από σελίδες που λήφθηκαν από το DBLP, όπως και από προσωπικές και ομαδικές ιστοσελίδες ερευνητών, ενώ το δεύτερο data set περιλαμβάνει 3.889 αναφορές από ταινίες του IMBD. Ο πίνακας 5 μας δείχνει την ευστοχία αντιστοίχισης του baseline αλγορίθμου χωρίς περιορισμούς στα data sets που περιγράψαμε (γραμμή 1), με χαλάρωση ετικετών (γραμμή 2) και τέλος με την πλήρη εφαρμογή του CME αλγορίθμου (γραμμή 3).

F1 (P / R)	Researchers	IMDB
Baseline	.66 (.67/.65)	.69 (.61/.79)
Baseline + Relax	.78 (.78/.78)	.72 (.63/.83)
Baseline + Relax + Pairwise	.79 (.80/.79)	.73 (.64/.83)

Πίνακας 5: ευστοχία αντιστοίχισης baseline αλγορίθμου

Αν M_a είναι το πλήθος όλων των σωστών αντιστοιχισμένων ζευγαριών και M_p το πλήθος των αντιστοιχισμένων ζευγαριών που επιστρέφει ο αλγόριθμος τότε θα έχουμε για την ακρίβεια $P = \frac{|M_p \cap M_a|}{|M_p|}$, την ανάκληση $R = \frac{|M_p \cap M_a|}{|M_a|}$ και τέλος για την $F-1 = \frac{(2PR)}{(P+R)}$.

Από τις τιμές του πίνακα 5 μπορούμε να παρατηρήσουμε το κέρδος που έχουμε από την εφαρμογή των περιορισμών, ενώ η προτεινόμενη μέθοδος απέδειξε πως μπορεί να βελτιώσει την τιμή του F1 ενός αλγορίθμου αντιστοιχίας οντοτήτων από 3% ως 12%.

2.5 Conclusions

Στο κεφάλαιο αυτό μελετήσαμε το πρόβλημα της εύρεσης του αντικειμένου που περιγράφει μια αναφορά και ειδικότερα αυτό μεταξύ paper – citation. Καταλήξαμε πως ποτέ δεν θα μπορέσουμε να επιτύχουμε μία απόλυτα σωστή αντιστοίχιση αφού φαινόμενα όπως το σφάλμα και η παρερμηνεία είναι πανταχού παρόν, αλλά με την βοήθεια αλγορίθμων συνοδευόμενοι από ορθούς περιορισμούς και μοντέλα πιθανοτήτων, μπορούμε να συμπεράνουμε αποτελέσματα τα οποία είναι πολύ κοντά στην αλήθεια.

3. Name Disambiguation

3.1 Introduction

Ακόμα και από την καθημερινή μας ζωή μπορεί να διαπιστωθεί πως ένα όνομα αντιστοιχίζεται σε περισσότερο από ένα φυσικά πρόσωπα αλλά και ένα πρόσωπο περιγράφεται με περισσότερο από ένα τρόπους. Μπορεί να υπάρχουν περισσότεροι από ένας Δημήτρης Παπαδόπουλος αλλά επίσης τα ονόματα Δημήτρης Παπαδόπουλος, Δημήτριος Παπαδόπουλος, Τάκης Π. να αναφέρονται στο ίδιο πρόσωπο. Η ασάφεια ονομάτων μπορεί να επηρεάσει την ποιότητα των συλλεγόμενων επιστημονικών δεδομένων, να μειώσει την απόδοση της ανάκτησης πληροφορίας και της διαδικτυακής αναζήτησης όπως και να προκαλέσει εσφαλμένα συμπεράσματα. Αυτό το κεφάλαιο ασχολείται με το πρόβλημα αποσαφήνισης φυσικών προσώπων και ειδικότερα με συντάκτες που εμφανίζονται στις διάφορες παραπομπές και έχουνε συμβάλει στην δημιουργία επιστημονικών άρθρων. Πρόκειται για ένα ιδιαίτερα σύνθετο αλλά και σημαντικό πρόβλημα, η αντιμετώπιση του οποίου θα μας επιτρέψει να βελτιώσουμε τα ποιοτικά αποτελέσματα αντιστοίχισης του κάθε επιστήμονα – συγγραφέα στα ανάλογα paper.

3.2 On co-authorship for author disambiguation

Κάθε συντάκτης αποτελεί μία οντότητα και μπορεί να περιγραφεί από τα διάφορα προσωπικά του στοιχεία όπως οι τίτλοι των papers που δημοσίευσε, η ημερομηνία γέννησής του, ο λογαριασμός email κτλ. Το συγκεκριμένο όμως άρθρο δίνει μεγάλη έμφαση στους συν-συγγραφείς του κάθε συγγραφέα και με βάση αυτή την πληροφορία προσπαθεί να αποσαφηνίσει τις ταυτότητες αυτών που μοιράζονται το ίδιο όνομα. Η λογική αυτής της προσέγγισης βασίζεται στο γεγονός ότι ο καθένας μας μπορεί να περιγραφεί από τα άτομα που γνωρίζει και συναναστρέφεται ενώ παράλληλα παρατηρείται πως γενικά οι συγγραφείς δεν τείνουν να αλλάζουν ιδιαίτερα συχνά τους συνεργάτες τους. Επομένως αν συναντήσουμε

ένα κοινό ζευγάρι ονομάτων σε δύο ή περισσότερες παραπομπές μπορούμε να υποθέσουμε με σχετική ασφάλεια πως αναφέρονται στα ίδια φυσικά πρόσωπα αλλά όχι πάντα λόγω ύπαρξης ασαφειών όπως η παρερμηνεία ονομάτων λόγω συντομογραφιών ή ακόμα και η απλή συνωνυμία.

Για παράδειγμα μας δίνονται οι παρακάτω πέντε παραπομπές σε καθεμία από τις οποίες εμφανίζεται το όνομα A. Cohen. Εξετάζοντας λίγο πιο προσεκτικά η C1 με την C2 όπως και η C3 με την C4 μοιράζονται κοινά ονόματα συν-συγγραφέων ενώ η C5 δεν έχει κανένα κοινό συν-συγγραφέα με τις υπόλοιπες παραπομπές. Αυτή η παρατήρηση μπορεί να μας οδηγήσει στο συμπέρασμα της ύπαρξης τριών διαφορετικών φυσικών προσώπων που μοιράζονται το ίδιο όνομα A. Cohen.

C1: A. Cohen, S. Draper, E. Martinian, G. Wornell (2006). Stealing bits from a quantized . . .
C2: A. Cohen, S. Draper, E. Martinian, G. Wornell (2002). Source requantization: successive . . .
C3: A. Cohen, J. Siegel, P. Rozin (2003). Faith versus practice: different bases for . . .
C4: A. Cohen, A. Malka, P. Rozin, L. Cherfas (2006). Religion and unforgivable offenses . . .
C5: A. Cohen, I. Tannenbaum (2001). Lesbian and bisexual women's judgments of . . .

Το χειρότερο σενάριο στο προηγούμενο παράδειγμα θα ήταν ένας συγγραφέας, πχ ο P. Rozin (C3,C4), να έχει συνεργαστεί με συγγραφείς που θα μοιράζονται το ίδιο όνομα, πχ A. Cohen. Αυτό το σενάριο αυξάνει την πολυπλοκότητα του προβλήματος και μπορεί να μας οδηγήσει σε λανθασμένα συμπεράσματα, πχ τέσσερις διαφορετικοί συγγραφείς αντί για τρείς που αναγνωρίζονται.

Μπορούμε να συμπεράνουμε πως μεταξύ διάφορων παραπομπών, όσους περισσότερους κοινούς συν-συγγραφείς έχει ένα όνομα συγγραφέα και όσο σπανιότερα είναι τα ονόματά τους τόσο το πιθανότερο το όνομα αυτό να αναφέρεται στο ίδιο πρόσωπο.

Είναι πιθανό κατά την αναζήτηση κοινών γνωστών συγγραφέων ενός κοινού ονόματος σε δύο διαφορετικές παραπομπές να καταλήξουμε με μηδενικά αποτελέσματα, αυτό όμως δεν καθιστά αναγκαίο πως τα δύο αυτά ονόματα αποτελούν και διαφορετικά πρόσωπά. Μία πιθανή μέθοδος αντιμετώπισης αυτού του προβλήματος θα ήταν να ελέγξουμε ένα επίπεδο βαθύτερα δηλαδή να λάβουμε υπόψη και τους γνωστούς των συν-συγγραφέων. Κάτι τέτοιο βέβαια θα μεγάλωνε κατά πολύ τις απαιτήσεις σε αποθηκευτικό χώρο και σε χρόνο εκτέλεσης του αλγορίθμου αλλά θα βελτίωνε την ποιότητα των αποτελεσμάτων.

3.2.1 Web-based acquisition of coauthors

Στην προσπάθειά μας να βρούμε όλους τους συν-συγγραφείς από το διαδίκτυο, με τη βοήθεια μηχανών αναζήτησης, θέτουμε ως ερώτηση κάθε φορά ένα ζευγάρι ονομάτων. Από τις επιστρεφόμενες ιστοσελίδες ελέγχουμε τα περιεχόμενα αυτών με την υψηλότερη κατάταξη, προκειμένου να βρούμε ονόματα συγγραφέων τα οποία θα θεωρηθούν ως αμοιβαία γνωστοί των δύο αυτών συγγραφέων. Η διαδικασία αυτή εκτελείται για κάθε ζευγάρι γνωστών συγγραφέων.

Δίνεται παρακάτω ο εν λόγω αλγόριθμος:

Αλγόριθμος 9: Web-based acquisition of coauthors
Input: a: the name of an author whose coauthors are to be gathered $C = \{c_1, \ldots, c_k\}$: a seed set of k known coauthors of a (k>=1) Initialize: $C_{new} \leftarrow C$ Loop: 1. $W_{new} \leftarrow 0$ 2. For each $c_i \in C_{new}$ 3. Search the Web for documents containing both last names of a and ci 4. Extract a set W of coauthors of a and ci from top-n retrieved documents 5. $W_{new} \leftarrow W_{new} \cup (W - C)$ 6. End For 7. Exit Loop if $W_{new} = 0$ 8. $C_{new} \leftarrow W_{new}$ 9. $C \leftarrow C \cup W_{new}$ Output: C: a set of expanded coauthors of a

Έστω μια αναφορά περιέχει τα ονόματα J. Smith, G. Patterson, A. Heisenberg, A. White. Για να βρούμε όλους τους γνωστούς συγγραφείς του J. Smith μπορούμε να χρησιμοποιήσουμε τον αλγόριθμο με δύο διαφορετικούς τρόπους. Στον πρώτο θέτουμε $a = 'J.Smith'$ και $C = \{'G.Patterson', 'A.Heisenberg', 'A.White'\}$ ως είσοδοι ενώ στον δεύτερο εκτελούμε επαναληπτικά τον αλγόριθμο αποκλείοντας τον συγγραφέα ‘J. Smith’ ως εξής, $< a = 'G.Patterson'$ και $C = \{'A.Heisenberg', 'A.White'\}>$, $< a = 'A.Heisenberg'$, $C = \{'G.Patterson', 'A.White'\}>$, $< a = 'A.White'$, $C = \{'G.Patterson', 'A.Heisenberg'\}>$, συγχωνεύοντας τα ονόματα που βρέθηκαν σε κάθε επανάληψη προκειμένου να πάρουμε το τελικό αποτέλεσμα.

3.2.2 Agglomerative Clustering for Same-name Author Occurrences

Αφού κάνουμε χρήση του προηγούμενου αλγορίθμου και συγκεντρώσουμε ένα πλήθος ονομάτων με τους συν-συγγραφείς του, στην συνέχεια θα θελήσουμε να αποσαφηνίσουμε το πλήθος των συγγραφέων που μοιράζονται κοινά ονόματα. Ο αλγόριθμος που παρουσιάζεται στη συνέχεια, δέχεται μια λίστα ίδιων ονομάτων συγγραφέων τα οποία εκπροσωπούνται από το πλήθος των συν-συγγραφέων τους. Αρχικά κάθε όνομα συγγραφέα θεωρείται ως ξεχωριστή οντότητα, στην συνέχεια όμως ο αλγόριθμος επαναληπτικά ελέγχει αν δύο οντότητες συγγραφέων είναι όμοιες εκτελώντας συγκρίσεις μεταξύ των συν-συγγραφέων τους. Αν το ποσοστό των κοινών γνωστών συγγραφέων ξεπερνά μία ορισμένη τιμή κατωφλίου θ τότε θεωρείται πως τα δύο ονόματα αναφέρονται στο ίδιο πρόσωπο και συγχωνεύονται.

Αλγόριθμος 10: Agglomerative Clustering for Same-name Author Occurrences

Input:

$a_1, \ldots, a_n$; same-name author occurrences

$a_i \leftarrow \{v_{i1}, \ldots, v_{im}\}$; each name occurrence a_i has a set of m (m>=0) his/her coauthor names

θ; a cluster-merging threshold

Initialize:

$c_i = \{a_i\}$; consider each name occurrence ai as an element of cluster ci

Loop:

1. DO
2. For each cluster-pair (c_i, c_j), calculate $CSim(c_i, c_j)$
3. $CSim(c_i, c_j) \leftarrow \max(ASim(a_x, a_y)), \forall a_x \in c_i, \forall a_y \in c_j$
4. $ASim(a_x, a_y) \leftarrow |a_x \cap a_y|$
5. Find the most similar cluster-pair (c_u, c_v)
6. $(c_u, c_v) \leftarrow \arg\max CSim(c_i, c_j)$
7. IF $CSim(c_u, c_v) \geq \theta$ THEN
8. $c_{u\_v} = c_u \cup c_v$; merge c_u and c_v into a new larger cluster c_{u_v}
9. ENDIF
10. WHILE $(CSim(c_u, c_v) \geq \theta)$

Output:

Clusters of author occurrences: $\{c_k\}$

3.2.3 Πειράματα

Αρχικά θα κάνουμε εισαγωγή σε έννοιες οι οποίες θα μας βοηθήσουν αργότερα στην εκτίμηση των πειραματικών αποτελεσμάτων.

$$precision = a/(a+b) \qquad recall = a/(a+c)$$

$$F1 = (presicion)(recall)/(precision + recall) \qquad accuracy = (a+d)/(a+b+c+d)$$

$$over-clustering\ \ error = b/(a+b+c+d)$$

$$under-clustering\ \ error = c/(a+b+c+d)$$

Η έννοιες των τιμών a, b, c, d δίνονται στον πίνακα 6.

		Σωστά ταιριασμένες ομάδες	
		ταιριασμένες	αταίριαστες
Παραγόμενες ομάδες	ταιριασμένες	a	b
	αταίριαστες	c	d

Πίνακας 6

Στο test set είχαμε κατά μέσο όρο 2,85 ρητούς συνεργάτες (EC), 19,75 έμμεσους συν-συγγραφείς συγγραφέων (aIC) και 17,65 έμμεσους συν-συγγραφείς συν-συγγραφέων (cIC), οι οποίοι προστέθηκαν από το web-based αλγόριθμο με τη μηχανή αναζήτησης της Google να ανακτά τα 20 σημαντικότερα κείμενα. Ο πίνακας 7 παρουσιάζει και συγκρίνει τα αποτελέσματα της ομαδοποίησης κάνοντας χρήση κάθε φορά διαφορετικά χαρακτηριστικά. Συγκρίνοντας τις δύο καλύτερες περιπτώσεις $EC + aIC$ και $EC + aIC + cIC$ παρατηρούμε πως η προσθήκη του cIC χαρακτηριστικού στην πρώτη περίπτωση ελάχιστα υποβαθμίζει τα αποτελέσματα, επειδή με το aIC δεν είναι συμπληρωματικά μεταξύ τους.

Features	Recall	Precision	F1	Under-clustering error	Over-clustering error
SC	0.7846	0.7017	0.7271	0.0000	0.3239
ST	0.2154	0.2063	0.2082	0.6761	0.0000
EC	0.8279	0.8725	0.8358	0.2303	0.0105
aIC	0.7770	0.7512	0.6994	0.0983	0.0722
cIC	0.6418	0.6308	0.5506	0.2498	0.0692
EC + aIC	0.8692	0.8820	0.8645	0.0900	0.0601
EC + cIC	0.8490	0.8743	0.8494	0.1478	0.0357
aIC + cIC	0.7901	0.7960	0.7550	0.1000	0.0873
EC + aIC + cIC	0.8693	0.8755	0.8609	0.0814	0.0848

Πίνακας 7: Επιδόσεις ομαδοποίησης με τον αριθμό επικάλυψης συν-συγγραφέων να ισούται με 1

3.3 Two Supervised Learning Approaches for Name Disambiguation in Author Citations

Στο παρόν paper γίνεται προσπάθεια αποσαφήνισης των συγγραφέων που καταγράφονται στις παραπομπές μέσω δύο επιβλεπόμενα μαθησιακών προσεγγίσεων. Η πρώτη προσέγγιση χρησιμοποιεί το γεννητικό στατιστικό μοντέλο Bayes, το οποίο αφού λάβει υπόψη πληροφορίες για κάθε συγγραφέα σχετικά με το επιστημονικό περιεχόμενο με το οποίο έχει ασχοληθεί αλλά και με τους συνεργάτες που έχει δουλέψει μαζί, εφαρμόζει πιθανοτικά μοντέλα προκειμένου να δημιουργήσει ένα πρότυπο γραφής για κάθε συγγραφέα. Το μοντέλο Bayes μπορεί να εκμεταλλευτεί μόνο τις ορθές παραπομπές για την μοντελοποίηση του μοτίβου του συγγραφέα, ενώ μπορεί να συνδυαστεί με άλλα μοντέλα [10] και να επεκταθεί κάνοντας χρήση επιπρόσθετες πληροφορίες. Η δεύτερη προσέγγιση εκμεταλλεύεται τις Support Vector Machines (SVM), η οποία διακρίνει τα δεδομένα μεταξύ τους με τη βοήθεια κλάσεων. Η μέθοδος αυτή θεωρεί κάθε συγγραφέα ως μία ξεχωριστή κλάση και ταξινομεί κάθε νέα παραπομπή στον πιθανότερο συγγραφέα. Επιπλέον μπορεί να αξιοποιήσει όλες τις παραπομπές ανεξαρτήτως εννοιολογικής ορθότητας ενώ τέλος χρησιμοποιεί μετρήσεις αποστάσεων[11] για την λήψη αποφάσεων από τιμές διανυσμάτων αναπαράστασης που περιγράφουν την παραπομπή τις οποίες συνήθως πρέπει να ορίσουμε[12,13]. Οι δύο προσεγγίσεις υποθέτουν την ύπαρξη μιας βάσης δεδομένων με παραπομπές και λαμβάνουν υπόψη τα εξής τρία χαρακτηριστικά:

Τα ονόματα των συν-συγγραφέων, τον τίτλο του paper και τους τίτλους των πηγών δημοσίευσής.

3.3.1 *The Naive Bayes Model*

Κάθε άνθρωπος είναι μοναδικός. Διαχειρίζεται καταστάσεις, αντιδρά σε προβλήματα και έχει ενδιαφέροντα διαφορετικά από οποιονδήποτε άλλο συνάνθρωπό του. Στο μοντέλο Bayes γίνεται προσπάθεια αξιοποίησης αυτού του φαινομένου, μοντελοποιώντας το προφίλ ενός συγγραφέα με υπολογισμό ανάλογων παραμέτρων, έχοντας πρώτα λάβει υπόψη γνωστές παραπομπές.

Βάση αυτών των παραμέτρων ο κανόνας του Bayes χρησιμοποιείται για να υπολογίσει την πιθανότητα με την οποία ένας συγγραφέας αναφέρεται σε μια παραπομπή έναντι άλλων υποψηφίων.

Όταν δεχόμαστε μία νέα παραπομπή C, θέλουμε να αποφασίσουμε ποιο από τα υπάρχοντα προφίλ συγγραφέων X_i είναι πιθανότερο να την έχει δημιουργήσει, δηλαδή $\max_i P(X_i \mid C)$.

Κάνοντας χρήση του μοντέλου Bayes ο προηγούμενος τύπος υπολογισμού πιθανότητας μετατρέπεται στον $\max_i P(C \mid X_i)P(X_i) / P(C)$.

Λαμβάνοντας υπόψη τα χαρακτηριστικά A_j των παραπομπών που προαναφέρθηκαν (A_1 - συν-συγγραφείς, A_2 - τίτλοι επιστημονικών άρθρων, A_3 - τίτλοι πηγών δημοσίευσής) και θεωρώντας τα ανεξάρτητα, καταλήγουμε στην εξίσωση $P(C \mid X_i) = \prod_j P(A_j \mid X_i) = \prod_j \prod_k P(A_{jk} \mid X_i)$, όπου K(j) το πλήθος των στοιχείων ενός χαρακτηριστικού A_j.

Για παράδειγμα θεωρώ $A_1 = (A_{11}, A_{12}, ..., A_{1k}, ..., A_{1k(1)})$, όπου A_{1k} ο k^{th} συγγραφέας.

Για να αποφευχθεί η περίπτωση υποχείλισης χρησιμοποιούμε λογαριθμικές τιμές πιθανοτήτων στην υλοποίησή μας, επομένως η συνάρτησή μας γίνεται $\max_i P(X_i \mid C) = \max_i \left[\sum_j \sum_k \log(P(A_{jk})) + \log(P(X_i)) \right]$, όπου $J \in [1,3]$, $k \in [0, K(j)]$. Η υπόθεση μας για ανεξαρτησία μεταξύ των χαρακτηριστικών A_j μπορεί να μην ισχύει πάντα για δεδομένα πραγματικού κόσμου, αλλά και με αυτή την παράβαση το μοντέλο Bayes παρουσιάζει καλή απόδοση.

Επειδή κάθε συγγραφέας έχει διαφορετική πιθανότητα να αναπτύξει ένα paper μόνος του, με γνωστούς ή νέους συν-συγγραφείς αλλά επίσης χαρακτηρίζεται και από το δικό του σύνολο γνωστών συγγραφέων, θα έχει και μοναδική πιθανότητα να συνεργαστεί με κάθε συγγραφέα ξεχωριστά. Θέλοντας να υπολογίσουμε την πιθανότητα $P(A_1 \mid X_i)$ με την οποία κάποιος συγγραφέας θα μπορούσε να συνεργαστεί με ένα πλήθος συγγραφέων, ορίζουμε και λαμβάνουμε υπόψη πιθανότητες οι οποίες εξατομικεύουν τον συγγραφέα αυτόν σαν οντότητα. Αντίστοιχα υπολογίζονται οι πιθανότητες $P(A_2 \mid X_i)$ και $P(A_3 \mid X_i)$ οι οποίες περιγράφουν κατά πόσο είναι πιθανό ο συγγραφέας X_i να ανέπτυξε άρθρο με ένα συγκεκριμένο τίτλο και να δημοσιεύτηκε σε κάποιο συγκεκριμένο χώρο αντίστοιχα.

- <u>Πολυπλοκότητα</u>

Εάν N ο αριθμός των συγγραφέων, M ο μέσος αριθμός παραπομπών που εμφανίζεται ένας συγγραφέας και K το πλήθος των χαρακτηριστικών, το μοντέλο Bayes υπολογίζει τις πιθανότητες σε $O(MNK)$ βήματα

3.3.2 *Support Vector Machines*

Κάθε συγγραφέας ταξινομείται σε μία κλάση με την βοήθεια των SVMs[14,15] και κάθε παραπομπή αντιπροσωπεύεται από ένα διάνυσμα χαρακτηριστικών. Τα ονόματα των συγγραφέων και οι λέξεις κλειδιά του τίτλου και του μέρους δημοσίευσης αποτελούν χαρακτηριστικά του εν λόγω διανύσματος, τα βάρη των οποίων εξαρτώνται από την συχνότητα εμφάνισής τους στην παραπομπή. Βάση αυτών των διανυσμάτων αποσαφηνίζεται σε ποιο πρόσωπο αναφέρεται ένα όνομα συγγραφέα. Η SVM μέθοδος έχει σχεδιαστεί να αντιμετωπίζει το πρόβλημα της ταξινόμησης, προσπαθώντας με βέλτιστο τρόπο να διαχωρίσει ένα σύνολο δεδομένων σε δύο κλάσεις. Έστω $\{(\vec{x}_1, y_1),....,(\vec{x}_N, y_N)\}$ όπου $\vec{x}_i$ ένα διάνυσμα χαρακτηριστικών και οι ετικέτες $y_i \subset (-1,+1)$. Η SVM προσπαθεί να ελαχιστοποιήσει την τιμή $\| \vec{w} \|$ έτσι ώστε $y_i(\vec{w} * \vec{x}_i + w_0) - 1 \geq 0, \forall i$ με την γραμμική συνάρτηση απόφασης να είναι η $f(\vec{x}) = \text{sgn}\{(\vec{w} * \vec{x} + w_0)\} = \text{sgn}\left\{\sum_i^n a_i^* y_i (\vec{x}_i * \vec{x}) + w_0^*\right\}$.

Αν η προηγούμενη συνάρτηση είναι θετική τότε τα δεδομένα $\vec{x}$ θα ανήκουν στην πρώτη κλάση διαφορετικά στην δεύτερη, με την απόλυτη τιμή του $f(\vec{x})$ να υποδηλώνει την απόσταση του $\vec{x}$ από την άλλη κλάση. Η μέθοδος μπορεί να επεκταθεί αν χρησιμοποιήσουμε την προσέγγιση μία κλάση εναντίον όλων των υπολοίπων.

3.3.3 **Πειράματα**

Μέσα από δύο πειράματα θα προσπαθήσουμε να μελετήσουμε τη συμπεριφορά των δύο μοντέλων ώστε να διαπιστώσουμε τις ιδιαιτερότητές τους και να αποφανθούμε τελικά για το ποιο από τα δύο είναι και αποτελεσματικότερο. Στο πρώτο πείραμα το dataset προέρχεται από καταλόγους δημοσιεύσεων που συλλέγονται από το διαδίκτυο, κυρίως από ιστοσελίδες ερευνητών. Στο δεύτερο πείραμα, το dataset έχει ληφθεί από την DBLP ιστοσελίδα και αποτελείται πάνω από 300.000 XML παραπομπές, στις οποίες τα ονόματα των συγγραφέων απλοποιούνται ("George Williams" – "G Williams") προκειμένου να βελτιώσουμε την ποιότητα του πειράματος.

$1^ο$ πείραμα:

Ελέγχουμε τα δύο μοντέλα σε δύο dataset από τις οποίες η πρώτη περιέχει 15 διαφορετικούς συγγραφείς "J Anderson" ενώ η δεύτερη 11 διαφορετικούς "J Smith". Αφού ληφθούν υπόψη τα κατάλληλα χαρακτηριστικά, θα γίνει προσπάθεια αποσαφήνισης των συγγραφέων από τα μοντέλα που μελετήθηκαν.

Scheme	Coauthor		Paper title		Journal title		Hybrid I	
Approach	**Bayes**	**SVM**	**Bayes**	**SVM**	**Bayes**	**SVM**	**Bayes**	**SVM**
Mean	71.3%	64.4%	77.9%	82.9%	72.1%	74.4%	91.3%	95.6%
StdDev	2.1%	3.8%	3.3%	1.9%	2.1%	3.0%	1.6%	1.7%
P Value	**1.38E-05**		**0.003**		**0.012**		**0.0003**	

πίνακας 8

Η μέση και τυπική απόκλιση της αποσαφήνισης των "J Anderson" (πίνακας 8) και "J Smith" (πίνακας 9) με την χρήση του Bayes και του SVM μοντέλου. Στο τέλος καταγράφεται η στατιστική σημαντικότητα της διαφοράς απόδοσης

Scheme	Coauthor		Paper title		Journal title		Hybrid I	
Approach	**Bayes**	**SVM**	**Bayes**	**SVM**	**Bayes**	**SVM**	**Bayes**	**SVM**
Mean	75.2%	60.0%	82.3%	84.2%	76.3%	78.4%	92.9%	94.5%
StdDev	3.0%	2.9%	3.5%	1.7%	2.2%	2.3%	2.0%	1.3%
P Value	**1.2E-09**		**0.074**		**0.035**		**0.031**	

πίνακας 9

Στους παραπάνω πίνακες εξετάζεται η συνεισφορά κάθε χαρακτηριστικού ξεχωριστά όπως και ο συνδυασμός αυτών (Hybrid I). Ο τίτλος του paper από μόνος του αποτελεί το πιο αποτελεσματικό χαρακτηριστικό έναντι των υπολοίπων ενώ όπως αναμενόταν, τα καλύτερα ποσοστά αποσαφήνισης διαπιστώθηκαν όταν λήφθηκαν υπόψη όλα τα χαρακτηριστικά, με ποσοστά πάνω από 90% για κάθε προσέγγιση και στις δύο περιπτώσεις, με το SVM μοντέλο να υπερισχύει αυτό του Bayes. Αξιοσημείωτη παρατήρηση αποτελεί η σημαντική υπερίσχυση του μοντέλου Bayes έναντι του SVM όσο αφορά τους συν-συγγραφείς και αυτό εξηγείται από το γεγονός ότι το δεύτερο δεν μπορεί να εκμεταλλευτεί συν-συγγραφείς που δεν έχει συναντήσει στο παρελθόν σε αντίθεση με το πρώτο που διαθέτει μηχανισμό πρόβλεψης βάση πιθανοτήτων.

$2^{ο}$ πείραμα:

Από τον πίνακα 10 παρατηρούμε πως οι δύο προσεγγίσεις παρουσιάζουν πολύ χειρότερη επίδοση συγκριτικά με τα αποτελέσματα του προηγούμενου πειράματος και γι αυτό ευθύνεται η χαμηλότερη ποιότητα των δεδομένων που χρησιμοποιούνται. Σε αυτή την περίπτωση οι συν-συγγραφείς συνεισφέρουν σημαντικά περισσότερο στην αποσαφήνιση από τα δύο υπόλοιπα χαρακτηριστικά με το μοντέλο του Bayes γενικά να υπερτερεί έναντι του SVM.

Scheme	Coauthor		Paper title		Journal title		Hybrid I	
Approach	**Bayes**	**SVM**	**Bayes**	**SVM**	**Bayes**	**SVM**	**Bayes**	**SVM**
Mean	69.3%	64.3%	18.9%	20.1%	40.0%	37.4%	69.1%	65.4%
StdDev	6.8%	8.3%	6.1%	6.9%	7.9%	7.6%	4.5%	3.8%
P Value	**0.010**		**0.497**		**0.053**		**0.009**	

πίνακας 10

- **<u>Πλεονεκτήματα</u>**

SVM: Καλή αξιολόγηση των χαρακτηριστικών μίας κλάσης

Bayes: Διορατικότερη προσέγγιση του προβλήματος, αποτελεσματικότερη εκμετάλλευση συν-συγγραφέων, ελαστικότερη επεξεργασία παραμέτρων

3.4 Name Disambiguation in Author Citations using a K-way Spectral Clustering Method

Η διαδικασία της αποσαφήνισης ονομάτων μπορεί να προσεγγιστεί με επιβλεπόμενες και μη-επιβλεπόμενες μεθόδους εκμάθησης. Στις επιβλεπόμενες μεθόδους κάθε όνομα συγγραφέα θεωρείται ως κλάση και κατά την διάρκεια αποσαφήνισης οι παραπομπές ταξινομούνται ανάλογα σε αυτές τις κλάσεις[16]. Για την ορθή εκτέλεση τους απαιτείται οι πληροφορίες να είναι επισημασμένες και να υπάρχουν ήδη καταγεγραμμένες αναγνωριστικές πληροφορίες για κάθε συγγραφέα ώστε η ταξινόμηση να γίνεται με σωστά κριτήρια. Αντιθέτως οι μη-επιβλεπόμενες μέθοδοι δεν απαιτούν την ύπαρξη επισημασμένων πληροφοριών με το πρόβλημα της αποσαφήνισης να διατυπώνεται ως ο διαχωρισμός των παραπομπών σε ομάδες, με κάθε ομάδα να περιέχει παραπομπές γραμμένες από τον ίδιο συγγραφέα. Ο πίνακας 11 παρουσιάζει ένα πλήθος παραπομπών τριών διαφορετικών συγγραφέων με το όνομα ‘J. E. Smith’ :

Ομάδα	Παραπομπές Συγγραφέων
1	Rapid Profiling via Stratified Sampling, S. Sastry, R. Bodik, **J. E. Smith**, 28th Int. Symposium on Computer Architecture, 2001.
	Relational Profiling: Enabling Thread-Level Parallelism in Virtual Machines, Timothy Heil and **J. E. Smith**, 33rd Int. Symp. on Microarchitecture, 2000.
	Concurrent Garbage Collection using Hardware Assisted Profiling, Timothy Heil and **J. E. Smith**, International Symposium on Memory Management, 2000.
2	**Smith, James E.**, “Moment Methods for Decision Analysis”, Management Science 39 (1993).
	Smith, James E., “Generalized Chebychev Inequalities: Theory and Applications in Decision Analysis”, Operations Research 43 (1995).
	Smith, James E., Samuel Holtzman and James E. Matheson, “Structuring Conditional Relationships in Influence Diagrams”, Operations Research 41 (1993).
	Henry E.J. and **Smith J.E.** 2002. The Effect of SurfaceActive Solutes on Water Flow and Contaminant transport in Variably Saturated Porous Media with Capillary Fringe Effects. Journal of Contaminant Hydrology.
	Henry E.J., **Smith J.E.**, and Warrick A.W. 2002.

3	Two-Dimensional Modeling of Flow and Transport in the Vadose Zone with Surfactant-Induced Flow. WATER RESOURCES RESEARCH.
	Smith, J.E. and Zhang F.Z. 2001. Determining Effective Interfacial Tension and Predicting Finger Spacing for DNAPL Penetration into Water-Saturated Porous Media. Journal of Contaminant Hydrology.

πίνακας 11: Παράδειγμα τριών διαφορετικών ομάδων αναφορών

Θα μελετήσουμε μία μη-επιβλεπόμενη μέθοδο αποσαφήνισης ονομάτων που θα χρησιμοποιεί το K-way φασματικό μοντέλο ομαδοποίησης [9], ένα μοντέλο γραφημάτων το οποίο εφαρμόζεται σε ζητήματα εξόρυξης πληροφορίας και ανάλυσης συστάδων. Αρχικά θα πρέπει να μοντελοποιηθεί κάθε παραπομπή ως κόμβος σε μη-κατευθυνόμενο γράφημα όπου σε κάθε ακμή (i, j) στον γράφο θα ανατίθεται μία τιμή που θα αντικατοπτρίζει την ομοιότητα μεταξύ των παραπομπών i και j. Γενικά οι φασματικές μέθοδοι ομαδοποίησης υπολογίζουν τις ιδιοτιμές και τα ιδιοδιανύσματα ενός πίνακα Laplace, σχετικό με ένα δοσμένο γράφο, και δημιουργούν ομάδες δεδομένων βασισμένες σε αυτή την φασματική πληροφορία [17,18,19,20,21]. Σύμφωνα με την [9], διαπιστώθηκε ότι η ελαχιστοποίηση του τετραγωνικού αθροίσματος μιας συνάρτησης κόστους μπορεί να αναδιατυπωθεί σε πρόβλημα μεγιστοποίησης ίχνους που σχετίζεται με τον πίνακα Gram των διανυσμάτων δεδομένων. Αποδεικνύεται ότι μια απλοποιημένη μορφή του πίνακα Gram μπορεί να μας δώσει βέλτιστες λύσεις για μια κανονικοποιημένη εκδοχή του προβλήματος μεγιστοποίησης ίχνους. Επομένως, η ανάθεση ομάδας για κάθε διάνυσμα δεδομένων μπορεί να βρεθεί υπολογίζοντας μία περιστρεφόμενη QR απλοποιημένη μορφή του πίνακα ιδιοδιανυσμάτων.

Για κάθε σύνολο ονομάτων, δημιουργούμε διάνυσμα παραπομπής M μεγέθους m, όσο δηλαδή του πλήθους των στοιχείων που λαμβάνουμε υπόψη. Ως στοιχείο θεωρούμε μία συνιστώσα ενός χαρακτηριστικού παραπομπής όπως το όνομα ενός συν-συγγραφέα ή μία λέξη από τον τίτλο του άρθρου κτλ. Δηλαδή στην σχέση $M = (a_1,...,a_m)$ το a_i αντιπροσωπεύει το βάρος του i στοιχείου. Αυτή η απόδοση τιμών μπορεί να γίνει από τα πρότυπα TFIDF και NTF με το δεύτερο να αποδεικνύει πως βελτιώνει τα αποτελέσματα ταξινόμησης [22].

$NTF(i,d) = \frac{freq(i,d)}{\max(freq(i,d))}$, όπου $freq(i,d)$ η συχνότητα εμφάνισης του στοιχείου i στην παραπομπή d.

Δοθέντος n διανύσματα παραπομπών m διαστάσεως το καθένα διαμορφώνουμε τον mn πίνακα $A=(a_1,...,a_n)$ και ορίζοντας ως E έναν πίνακα μετάθεσης μπορούμε να γράψουμε μία διαμέριση Π από διανύσματα παραπομπών στην μορφή

$AE=[A_1,...,A_k], A_i=[a_1^{(i)},...,a_{si}^{(i)}]$.

Δοθέντος μια διαμέριση Π η συνάρτηση κόστους τετραγωνικού αθροίσματος είναι η εξής

$$ss(\Pi)=\sum_{i=1}^{k}\sum_{s=1}^{si}\| a_s^{(i)}-m_i\|^2, m_i=\sum_{s=1}^{si} a_s^{(i)} \Big/ s_i$$

Από την [52] διαπιστώνεται πως το παραπάνω άθροισμα μπορεί να διατυπωθεί ως ένα κανονικοποιημένο πρόβλημα μεγιστοποίησης ως εξής:

$\max[trace(X^T A^T AX)]$, όπου $X^T X=I_K$ και X ένα τυχαίος ορθοκανονικός πίνακας.

<u>Θεώρημα:</u> Για κάθε συμμετρικό πίνακα H με ιδιοτιμές $\lambda_1\geq\lambda_2\geq....\geq\lambda_n$ και με αντίστοιχα ιδιοδιανύσματα $U=[u_1,...,u_n]$ θα έχω $\lambda_1+...+\lambda_n=\max_{X^TX=I_K} trace(X^THX)$

Με βάση το προηγούμενο θεώρημα θα πρέπει να υπολογίσουμε τα k μεγαλύτερα ιδιοδιανύσματα από έναν Gram πίνακα A^TA προκειμένου να λάβουμε την δομή κάθε ομάδας παραπομπών. Έστω X_k ο nk πίνακας που περιέχει τα ζητούμενα ιδιοδιανύσματα με κάθε του γραμμή να αντιστοιχίζεται με ένα διάνυσμα παραπομπής.

Έστω $A=[A_1,...,A_k]$ η βέλτιστη διαμέριση από διανύσματα παραπομπών, που ελαχιστοποιεί την τιμή της $ss(\Pi)$. Ο πίνακας Gram του A μπορεί να γραφτεί ως

$$A^TA=\begin{pmatrix} A_1^TA_1 & 0 & . & 0 \\ 0 & A_2^TA_2 & . & 0 \\ 0 & 0 & . & A_k^TA_k \end{pmatrix}+E\equiv B+E$$

Όταν η επικάλυψη μεταξύ των ομάδων A_i είναι μικρή, η νόρμα του E θα είναι αρκετά μικρότερη σε σχέση με αυτή του πίνακα B. Αν y_i το μεγαλύτερο ιδιοδιάνυσμα του πίνακα $A_i^TA_i$ και $A_i^TA_i\,y_i=u_iy_i, \|y_i\|=1, i=1,...,k$ τότε οι στήλες του πίνακα

$$Y_k=\begin{matrix} s_1 \\ s_2 \\ \vdots \\ s_k \end{matrix}\begin{pmatrix} y_1 & & \\ & y_2 & \\ & & \ddots \quad y_k \end{pmatrix}$$ εκτείνονται σε ένα αμετάβλητο υποχώρο του B.

Λαμβάνοντας υπόψη τα ιδιοδιανύσματα και τις ιδιοτιμές του πίνακα Gram καταλήγουμε στην σχέση $X_k^T\equiv[x_1,...,x_k]=Y_kV+O(\|E\|)$, V ένας kk ορθογώνιος πίνακας. Εάν αγνοήσουμε τον

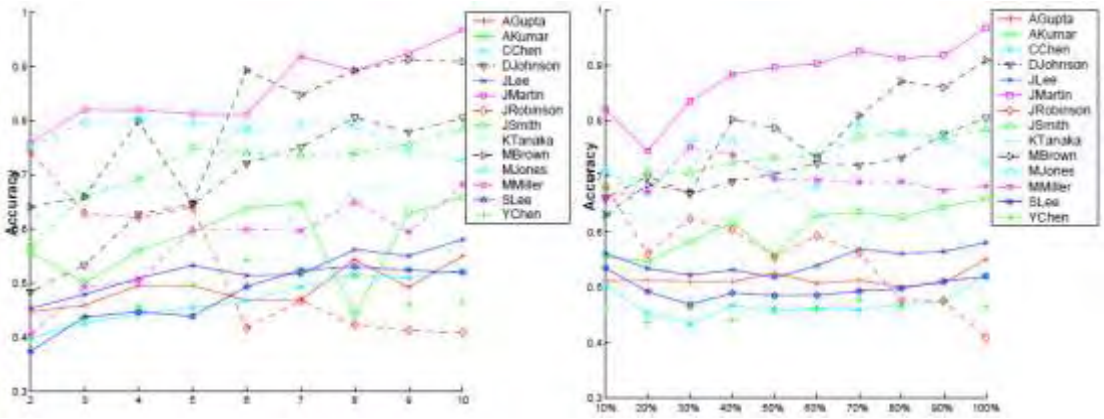
όρο $O(\|E\|)$ παρατηρούμε πως $X_k^T = \left[\underbrace{y_{11}u_1,...,y_{1s_1}u_1}_{cluster1},..,\underbrace{y_{k1}u_k,...,y_{ks_k}u_k}_{clusterk} \right]$, όπου u_i ορθογώνια μεταξύ τους. Αν επιλέξουμε μία u_i στήλη μπορούμε να μεταβούμε σε άλλες ομάδες εξετάζοντας τα ορθογώνια της συμπληρώματα. Επιλέγοντας δηλαδή την στήλη του X_K^T με την μεγαλύτερη νόρμα, έστω ότι ανήκει στην i ομάδα, ελέγχουμε το εσωτερικό γινόμενο της με κάθε άλλη στήλη. Αυτές που το αποτέλεσμα της πράξης είναι σχετικά μικρό θεωρούμε πως ανήκουν στην ίδια συστάδα. Στην συνέχεια ο αλγόριθμος εκτελεί ακριβώς την ίδια διαδικασία για τις επόμενες $k-1$ μεγαλύτερες σε νόρμα στήλες. Αποδεικνύεται πως εκτελούνται ακριβώς QR απλοποιήσεις με περιστροφή στηλών που εφαρμόζονται στον X_K^T ως εξής $X_K^T P = QR = Q[R_{11}, R_{12}]$, όπου P πίνακας περιστροφής, Q ορθογώνιος πίνακας kk και R_{11} άνω τριγωνικός πίνακας kk.

Τέλος υπολογίζουμε τον πίνακα $\hat{R} = R_{11}^{-1}[R_{11}, R_{12}]P^T = [I_k, R_{11}^{-1}R_{12}]P^T$, όπου τα μέλη ομάδας κάθε διανύσματος παραπομπής καθορίζονται από τον δείκτη γραμμής του μεγαλύτερου στοιχείου σε απόλυτη τιμή της αντίστοιχης στήλης του $\hat{R}$.

- **<u>Πειράματα</u>**

Τα πειράματα που διεξήχθησαν έχουν εφαρμοστεί σε δύο διαφορετικά τύπου datasets. Το πρώτο περιέχει 400.000 αναφορές σε XML μορφή κατεβασμένες από την DBLP ιστοσελίδα ενώ το δεύτερο ανακτήθηκε χειροκίνητα από το ιστοσελίδες ερευνητών με την βοήθεια μηχανών αναζήτησης. Μελετήθηκε και αξιολογήθηκε η συνεισφορά που μπορεί να έχουν οι συν-συγγραφείς, ο τίτλος του άρθρου και ό τόπος δημοσίευσης του ξεχωριστά, ενώ για κάθε σύνολο ονομάτων, έχει ληφθεί υπόψη η επιρροή του πλήθους των παραπομπών που συσχετίζεται με αυτά, όπως επίσης και τα πρότυπα απόδοσης τιμών TFIDF και NTF στα βάρη των στοιχείων.

3.4.1 Η εικόνα 10 παρουσιάζει σχηματικά, ενώ ο πίνακας 12 ποσοτικά, την ακρίβεια αποσαφήνισης του K-way φασματικού μοντέλου ομαδοποίησης όσο αφορά το πλήθος των αναφορών που λαμβάνουμε υπόψη (από 10% έως 100% με βήμα 10%) για ορισμένα σύνολα ονομάτων συγγραφέων.



Εικόνα 10: Στο αριστερό σχήμα εμφανίζονται τα αποτελέσματα από το πρώτο data set ενώ στο δεξί σχήμα το δεύτερο. Η απόδοση βάρους γίνεται από το TFIDF πρότυπο.

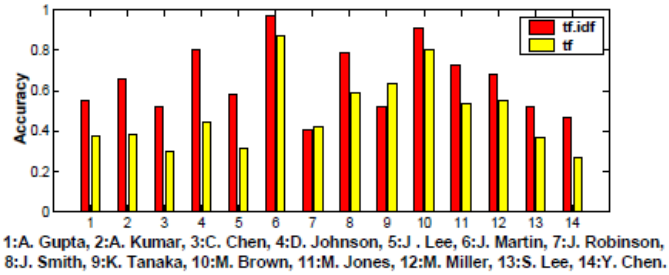
Name	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
A. Gupta	51.1	51.3	51.0	51.0	52.5	50.7	51.2	50.1	50.8	53.9
A. Kumar	55.6	54.6	58.2	61.6	55.7	63.0	63.4	62.6	64.5	64.3
C. Chen	50.4	45.2	43.3	46.8	45.6	46.3	45.8	47.0	47.3	50.6
D. Johnson	66.0	70.0	66.9	69.1	70.4	72.4	72.0	73.3	77.6	79.1
J. Lee	56.1	53.4	52.2	53.1	51.7	53.8	56.9	56.0	56.4	56.2
J. Martin	82.0	74.5	73.6	88.3	89.7	90.3	92.6	91.1	91.8	96.8
J. Robinson	68.2	56.0	62.4	60.4	55.4	59.3	56.2	47.6	47.5	39.2
J. Smith	67.7	70.8	70.6	72.5	73.3	73.7	77.2	77.9	77.0	77.4
K. Tanaka	63.9	62.7	62.4	62.1	67.7	63.9	56.7	69.8	62.5	50.8
M. Brown	63.0	68.5	67.1	80.2	78.7	73.3	81.0	87.1	86.0	87.0
M. Jones	71.0	66.8	76.3	76.4	70.0	68.0	79.4	77.3	76.6	70.6
M. Miller	66.5	67.2	75.2	74.0	69.5	69.3	68.9	69.1	67.4	67.4
S. Lee	53.4	49.2	46.9	48.9	48.5	48.5	49.3	49.8	51.1	50.4
Y. Chen	46.1	43.6	46.1	44.2	46.2	45.9	47.8	46.4	47.3	45.5
Mean	61.5	59.6	60.9	63.5	62.5	62.7	64.2	64.7	64.6	63.5
Std	9.4	9.9	11.0	13.2	13.0	12.4	14.0	14.9	14.7	16.8

πίνακας 12: Ευστοχία αποσαφήνισης 14 ονομάτων από το DBLP dataset

Σε γενικές γραμμές παρατηρούμε πως όσο περισσότερες παραπομπές λαμβάνουμε υπόψη τόσο αυξάνεται η ευστοχία της αποσαφήνισης, αλλά αυτό δεν ισχύει πάντοτε. Μπορεί δύο επιστήμονες που μοιράζονται το ίδιο όνομα να έχουν κοινό ερευνητικό πεδίο ίσως και συν-συγγραφείς με κοινά ονόματα με αποτέλεσμα τα αποτελέσματα να φθίνουν σε ακρίβεια καθώς το πλήθος των παραπομπών αυξάνεται. Ένας τρόπος αντιμετώπισης θα ήταν να αυξήσουμε τα χαρακτηριστικά που θα λαμβάνουμε υπόψη κατά την αποσαφήνιση.

3.4.2 Όσο αφορά τα πρότυπα απόδοσης τιμών, η χρήση του TFIDF μας επέστρεψε πολύ πιο ακριβή αποτελέσματα σε σχέση με αυτά του NTF και ο λόγος είναι πως το πρώτο πρότυπο πέρα από την συχνότητα εμφάνισης ενός στοιχείου, λαμβάνει επιπλέον υπόψη την διανομή του σε όλες τις παραπομπές που σχετίζονται με ένα όνομα. Στο γράφημα την εικόνας 11

εφαρμόζεται σε 14 ονόματα το TFIDF (αριστερά) και NTF (δεξιά) πρότυπο και εξετάζουμε την ακρίβεια αποσαφήνισης.



Εικόνα 11: Σύγκριση TFIDF και NTF

3.4.3 Θέλοντας να αποφασίσουμε ποιο από τα χαρακτηριστικά μιας παραπομπής συνεισφέρει περισσότερο στην σωστή αποσαφήνιση των κοινών ονομάτων, εκτελέστηκε πείραμα στο οποίο λήφθηκε υπόψη κάθε ένα από αυτά ξεχωριστά. Στον πίνακα 13, η στήλη coauthor 1 θεωρεί πως τα ονόματα που δεν έχει συν-συγγραφείς αποσαφηνίσθηκαν λανθασμένα ενώ η στήλη coauthor 2 δεν εφαρμόζει κάποιον περιορισμό.

Name	Coauthor 1	Coauthor 2	PTitle	Venue title
A. Gupta	37.9%	39.8%	47.7%	24.7%
A. Kumar	25.7%	34.0%	61.0%	45.2%
C. Chen	33.3%	37.3%	43.7%	23.7%
D. Johnson	31.9%	41.2%	53.4%	50.0%
J. Lee	38.8%	45.1%	38.1%	19.6%
J. Martin	37.9%	62.5%	50.0%	65.2%
J. Robinson	41.2%	53.0%	43.2%	37.2%
J. Smith	46.7%	58.4%	44.0%	24.7%
K. Tanaka	49.6%	54.5%	68.6%	46.5%
M. Brown	50.4%	57.4%	61.7%	36.5%
M. Jones	43.9%	61.8%	50.2%	33.5%
M. Miller	52.4%	53.7%	52.4%	53.0%
S. Lee	34.3%	36.1%	37.7%	30.4%
Y. Chen	37.3%	43.1%	31.2%	19.8%
Mean	40.1%	48.4%	48.8%	36.4%
Std	7.7%	10.0%	10.3%	13.9%

πίνακας 13: Συνεισφορά των metadata

Παρατηρούμε πως χρησιμοποιώντας μόνο τους συν-συγγραφείς, έχουμε περίπου παρόμοια ακρίβεια με αυτή του τίτλου του paper. Και οι δύο προηγούμενες περιπτώσεις ξεπερνούν την ακρίβεια που μας προσφέρει ο τίτλος του τόπου δημοσίευσης.

3.5 Conclusions

Σε αυτό το κεφάλαιο μελετήσαμε μεθόδους και τεχνικές με τις οποίες μπορούμε να επιτύχουμε αποσαφήνιση μεταξύ συγγραφέων που μοιράζονται το ίδιο όνομα. Το πρόβλημα αυτό όπως και κάθε πρόβλημα ασάφειας δεν μπορεί να προσεγγιστεί με κάποιον απόλυτα σωστό τρόπο όμως στα paper αυτά προσφέρονται αποδεκτές λύσεις που επιστρέφουν αρκετά ρεαλιστικά αποτελέσματα.

4. References

[1] Bing Liu, Robert Grossman, Yanhong Zhai. Mining data records in web pages.

[2] Steve Lawrence, C. Lee Giles, Kurt D. Bollacker. Autonomous citation Matching.
[3] Hanna Pasula, Bhaskara Marthi, Brian Milch, Stuart Russell, Ilya Shpitser. Identity uncertainty and citation matching.
[4] Warren Shen, Xin Li, AnHai Doan. Constraint-based entity matching.
[5] In-Su Kang, Seung-Hoon Na, Seungwoo Lee, Hanmin Jung, Pyung Kim, Won-Kyung Sung, Jong-Hyeok Lee. On co-authorship for author disambiguation.
[6] Hui Han, Lee Giles, Hongyuan Zha, Cheng Li, Kostas Tsioutsiouliklis. Two supervised learning approaches for name disambiguation in author citations
[7]Hui Han, Hongyan Zha, C. Lee Giles. Name disamdiguation in Author Citations using a K-way spectral clustering method.
[8] William W. Cohen, Pradeep Ravikumar, Stephen E. Fienberg. A comparison of string distance metrics for name-matching tasks
[9] H. Zha, C. Ding, M. Gu, X. He, and H. Simon. Spectral relaxation for k-means clustering. In *Neural Information Processing Systems (NIPS 2001)*, pages 1057–1064, 2001.
[10] T. Hofmann. Probabilistic latent semantic analysis. In *Proceedings of Uncertainty in Artificial Intelligence, UAI'99*, 1999.
[11] K. Takeuchi and N. Collier. Use of support vector machines in extended named entity, 2002.
[12] M. Bilenko, R. Mooney, W. Cohen, P. Ravikumar, and S. Fienberg. Adaptive name matching in information integration. *IEEE Intelligent Systems*, 18(5):16–23, 2003.
[13] S. Tejada, C. Knoblock, and S. Minton. Learning domain-independent string transformation weights for high accuracy object identification. In *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 350–359, 2002.
[14] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, 1995.
[15] N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines*. Cambridge University Press, 2000.
[16] H. Han, C. L. Giles, H. Zha, C. Li, and K. Tsioutsiouliklis. Two supervised learning approaches for name disambiguation in author citations. In *Proceedings of the 4th ACM/IEEE-CS Joint Conference on Digital libraries*, 2004.
[17] P. Drineas, A. Frieze, R. Kannan, S. Vempala, and V. Vinay. Clustering in large graphs and matrices. In *SODA: ACM-SIAM Symposium on Discrete Algorithms*, pages 291–299, 1999.
[18] R. Kannan, S. Vempala, and A. Vetta. On clusterings: Good, bad and spectral. In *Proceedings of the 41st Foundations of Computer Science*, pages 367–380, 2000.
[19] A. Ng, M. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. In *Proceedings of Advances in Neural Information Processing Systems*, pages 849–856, 2001.
[20] A. Pothen, H. D. Simon, and K.-P. Liou. Partitioning sparse matrices with eigenvectors of graphs. 11:430–452, 1990.
[21] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000.
[22] H. Han, C. L. Giles, E. Manavoglu, H. Zha, Z. Zhang, and E. A. Fox. Automatic document metadata extraction using support vector machines. In *Proceedings of the 3rd ACM/IEEE-CS Joint Conference on Digital libraries*, pages 37–48, 2003.
[23] Baeza-Yates, R. .Algorithms for string matching: A survey.. *ACM SIGIR Forum*, 23(3-4):34--58, 1989
[24] Gusfield, D. *Algorithms on strings, tree, and sequence*. 1997.
[25] Buttler, D., Liu, L., Pu, C. "A fully automated extraction system for the World Wide Web." *IEEE ICDCS-21*, 2001.
[26] Chang, C-H., Lui, S-L. .IEPAD: Information extraction based on pattern discovery.. *WWW-10*, 2001.
[27] A. Pfeffer. *Probabilistic Reasoning for Complex Systems*. PhD thesis, Stanford, 2000.
[28] A. Pfeffer and D. Koller. Semantics and inference for recursive probability models. In *AAAI/IAAI*, 2000.
[29] W.R. Gilks, S. Richardson, and D.J. Spiegelhalter. *Markov chain Monte Carlo in practice*. Chapman and Hall, London, 1996.
[30] Monge, A., and Elkan, C. 1996. The field-matching problem: algorithm and applications. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*.
[31] Jaro, M. A. 1989. Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida. *Journal of the American Statistical Association* 84:414–420.
[32] Jaro, M. A. 1995. Probabilistic linkage of large public health data files (disc: P687-689). *Statistics in Medicine* 14:491–498.
[33] Winkler, W. E. 1999. The state of record linkage and current research problems. Statistics of Income Division, Internal Revenue Service Publication R99/04. Available from http://www.census.gov/srd/www/byname.html.
[34] Fellegi, I. P., and Sunter, A. B. 1969. A theory for record linkage. *Journal of the American Statistical Society* 64:1183–1210.
[35] Monge, A., and Elkan, C. 1996. The field-matching problem: algorithm and applications. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*.
[36] Monge, A., and Elkan, C. 1997. An efficient domain-independent algorithm for detecting approximately duplicate database records. In *The proceedings of the SIGMOD 1997 workshop on data mining and knowledge discovery*.