HY463 - Συστήματα Ανάκτησης Πληροφοριών
Information Retrieval (IR) Systems

Parallel and Distributed IR
Παράλληλη και Κατανεμημένη ΑΠ

Ενοποίηση Αποτελεσμάτων
(… Results Merging, Fusion, Rank Aggregation, ...)

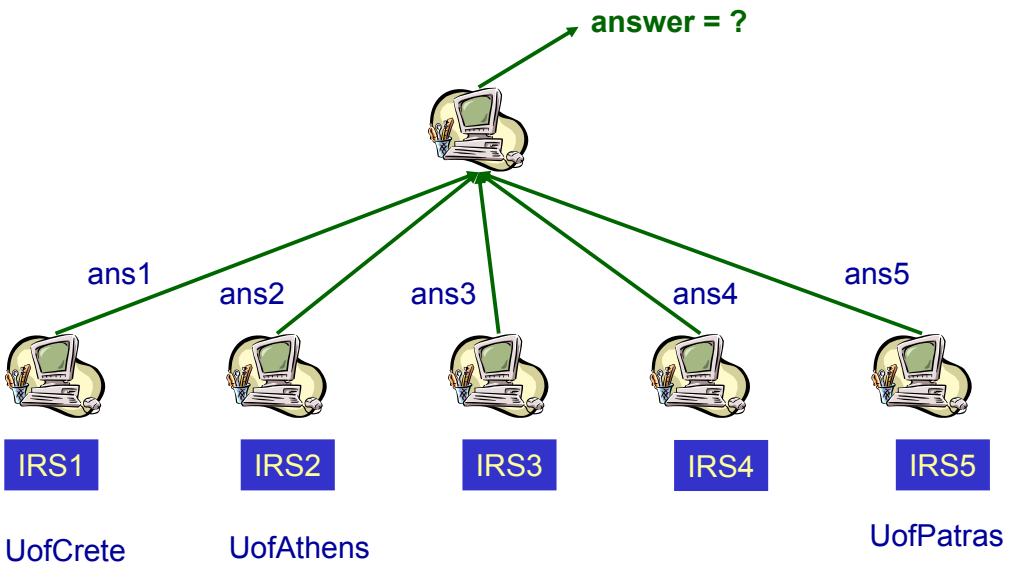
Γιάννης Τζίτζικας

Διάλεξη : 17
Ημερομηνία : 30-5-2007

Ενοποίηση Αποτελεσμάτων
Διάρθρωση

- Κατηγορίες Τεχνικών Ενοποίησης: Isolated vs Integrated
- Τεχνικές Ενοποίησης
  - Round Robin interleaving
  - Score-based
  - Weighted Score-based
  - Global-statistics
- Μετα-Μηχανές Αναζήτησης
- Ενοποίηση Διατάξεων (Rank-Aggregation)
  - Επιθυμητές Ιδιότητες
  - Ενοποίηση Borda
  - Ενοποίηση Condorcet
  - Το Θεώρημα του Ανέφικτου του Arrow (Arrow's Impossibility theorem)
  - Ενοποίηση Kemeny

Ενοποίηση Αποτελεσμάτων



Περιπτώσεις

- Ενοποίηση **Συνόλων** (π.χ. απαντήσεων σε Exact Match Queries)
  - answer(q) = ans1(q) ∪ … ∪ ansk(q)
  - Άρα η ενοποίηση αποτελεσμάτων για το Boolean model είναι εύκολη

- Ενοποίηση **Διατάξεων** (απαντήσεων Partial Match Queries)
  - Η ενοποίηση αποτελεσμάτων είναι πιο δύσκολη
  - οι διατάξεις/σκορ **δεν είναι πάντα συγκρίσημες** (αφού εξαρτώνται από τα στατιστικά της συλλογής του κάθε συστήματος (e.g. idf)

Κατηγορίες Στρατηγικών Διατάξεων

(A) **Ολοκληρωμένες** Τεχνικές (**Integrated**)

- Οι πηγές παρέχουν **επιπρόσθετη πληροφορία** που χρησιμοποιείται κατά την ενοποίηση
- Αδυναμίες:
  - Στενό πεδίο εφαρμογής - απαιτούν συμφωνία μεταξύ των πηγών (e.g. protocol)
  - Συχνά λαμβάνουν υπόψη τους μέτρα όπως Precision/Recall, τα οποία δεν είναι αντικειμενικά ή συγκρίσιμα.

(B) **Απομονωμένες** Μέθοδοι (**Isolated**)

- Δεν απαιτούν **καμία επιπλέον πληροφορία** από τις πηγές (μπορούν να εφαρμοστούν και στις μετα-μηχανές αναζήτησης)
- Είναι ανεξάρτητες των τεχνικών ευρετηρίασης και των μοντέλων ανάκτησης των υποκείμενων συστημάτων
- Άρα κατάλληλες για δυναμικά περιβάλλοντα όπου υπάρχουν πολλά συστήματα των οποίων η λειτουργία εξελίσσεται συχνά και απρόβλεπτα
- Τεχνικές:round robin interleaving, score-based, Borda, Condorcet, download and re-index the contents of the objects (web pages)

Ενοποίηση Διατάξεων: **Round Robin interleaving** (isolated)

(δηλαδή merge sort)

- Παράδειγμα:
  - ans1(q) = <d10,d2, d30, d7>
  - ans2(q) = <d4, d12, d5, d9>
  - ANS(q) = < {d10,d4}, {d2,d12}, {d30,d5}, {d7,d9}>
- Προβλήματα
  - στην πραγματικότητα όλα τα έγγραφα του ans1(q) μπορεί να είναι καλύτερα (πιο συναφή) από το 1ο στοιχείο της ans2(q)

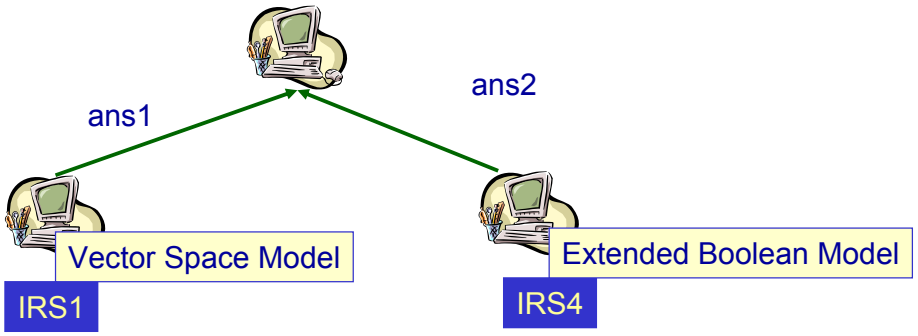
Ενοποίηση Διατάξεων: **Score-based** (isolated)

- Παράδειγμα:
  - ans1(q) = < (d3,0.8), (d2,0.7) >
  - ans2(q) = < (d5,0.6), (d6,0.3) >
  - ans3(q) = < (d4,0.9) >
  - ANS(q) = < d4, d3, d2, d5, d6>
- Προβλήματα
  - τα σκορ διαφορετικών συστημάτων δεν είναι συγκρίσιμα (κανονικοποιημένα), αφού εξαρτώνται από τα στατιστικά της συλλογής του κάθε συστήματος (e.g. idf)

Ενοποίηση Διατάξεων: **Weighted Score-based**

- Λαμβάνουμε υπόψη το σκορ της πηγής που υπολογίσαμε όταν κάναμε *Επιλογή Πηγής*
  - Πχ
  - Sc(IRS1) = 0.9 // υπολογίστηκε στη φάση επιλογή πηγής
  - Sc(IRS2) = 0.5 // υπολογίστηκε στη φάση επιλογή πηγής
  - ans1(q) = <(d1, 0.7)>
  - ans2(q) = <(d2, 0.9)>
  - ANS(q) = < (d1, 0.56), (d2, 0.45)> // 0.63 = 0.9*0.7
- Εδώ πολλαπλασιάσαμε το σκορ της πηγής με το σκορ των εγγράφων. Διάφορες άλλες παραλλαγές υπάρχουν (Callan94,95)

Ενοποίηση Διατάξεων: **Download and re-index/re-score** (isolated)



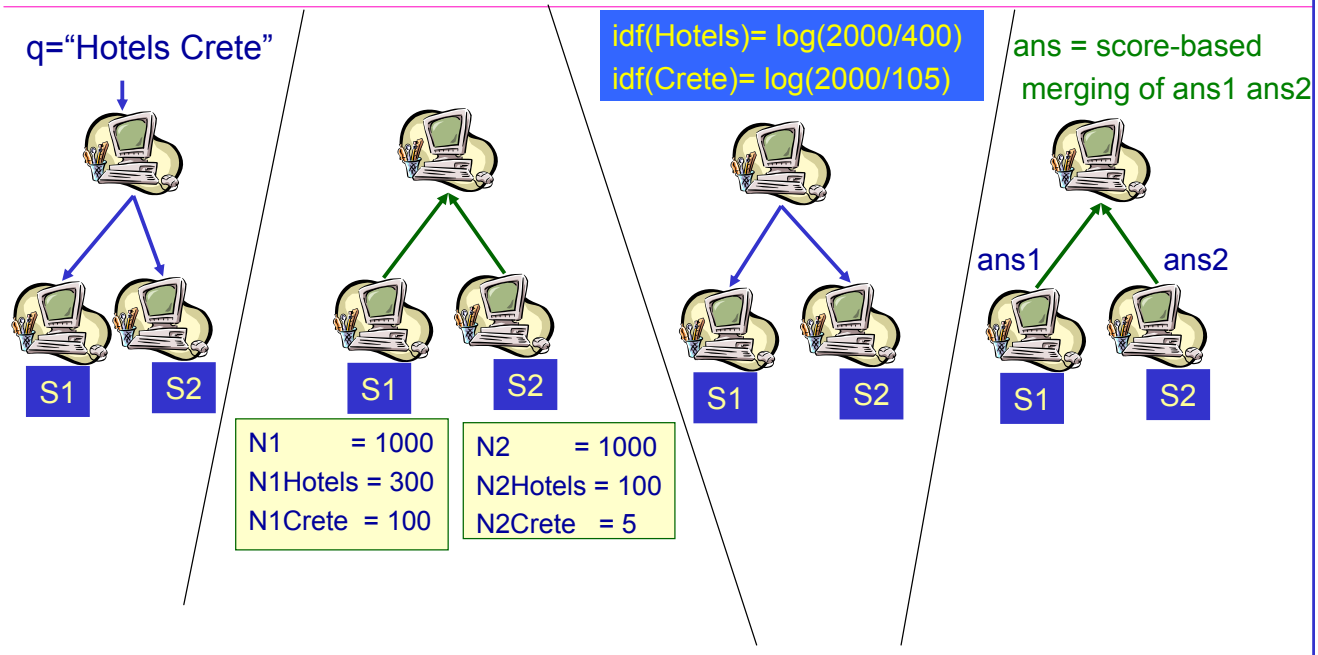
- Ανακτούμε τα έγγραφα των απαντήσεων κάθε πηγής
- Τα επαναευρετηριάζουμε και ξαναυπολογίζουμε το βαθμό συνάφειας τους
- Αδυναμίες
  - Χρονοβόρα διαδικασία

Ενοποίηση Διατάξεων: **Global term statistics (integrated)**

- Μπορούμε να κάνουμε συγκρίσιμα τα σκορ διαφορετικών συστημάτων αν επιβάλουμε τα ίδια στατιστικά στοιχεία σε όλα τα συστήματα (global statistics)

- Τρόποι απόκτησης αυτών των στοιχείων
  - Κατά την επιλογή πηγής (πχ Διανύσματα Πηγής, Probe Queries, …)
  - Αποτίμηση Επερωτήσεων σε 2 φάσεις
    - στην 1η συλλέγονται τα στατιστικά (ο server στέλνει την επερώτηση και οι πηγές απαντούν με τα στατιστικά των όρων που περιέχονται στην επερώτηση)
    - στην 2η ο server στέλνει σε κάθε πηγή την επερώτηση μαζί με τα καθολικά στατιστικά των όρων της
    - κάθε πηγή αποτιμά την επερώτηση με τα καθολικά στατιστικά και επιστρέφει την απάντηση
    - Ο server λαμβάνει έτοιμα σκορ και απλά τα ενοποιεί (merge sort)

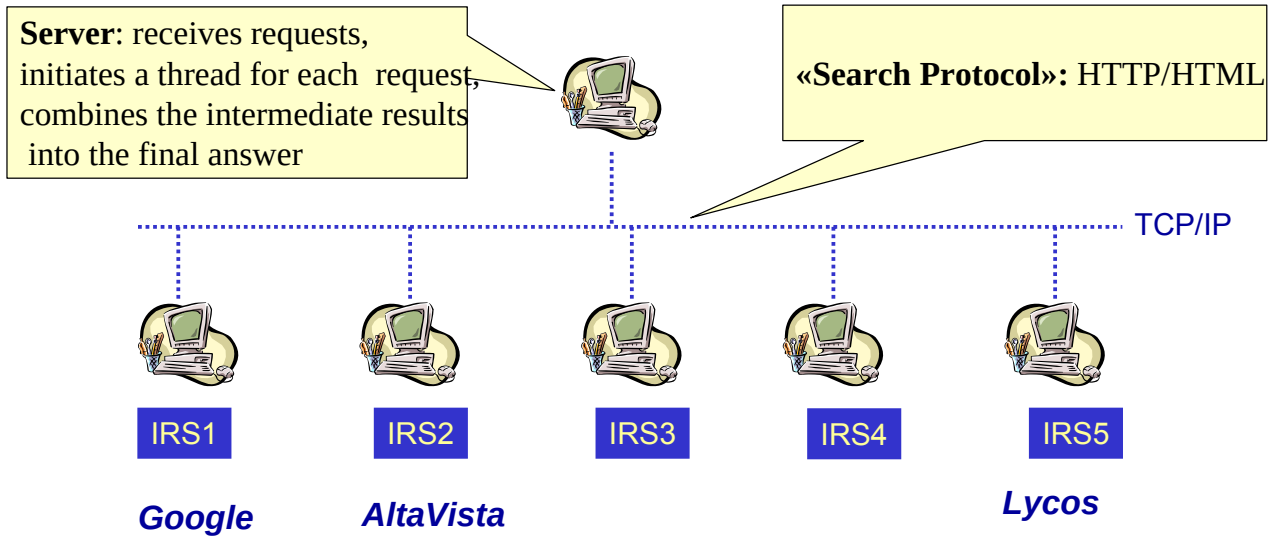
Ενοποίηση Διατάξεων: Global term statistics Παράδειγμα



Ενοποίηση Αποτελεσμάτων Διάρθρωση

- Κατηγορίες Τεχνικών Ενοποίησης: Isolated vs Integrated
- Τεχνικές Ενοποίησης
  - Round Robin interleaving (isolated)
  - Score-based (isolated)
  - Weighted Score-based (integrated)
  - Global-statistics (integrated)
- Μετα-Μηχανές Αναζήτησης
- Ενοποίηση Διατάξεων (Rank-Aggregation)
  - Επιθυμητές Ιδιότητες
  - Ενοποίηση Borda
  - Ενοποίηση Condorcet
  - Ενοποίηση Kemeny
  - Το Θεώρημα του Ανέφικτου του Arrow (Arrow's Imposibility theorem)

Μετα-Μηχανές Αναζήτησης



Μετα-Μηχανή Αναζήτησης: Μηχανή αναζήτησης που προωθεί την επερώτηση σε πολλές μηχανές αναζήτησης και ενοποιεί τα αποτελέσματα που επιστρέφουν

Γιατί φτιάχνουμε μετα-μηχανές αναζήτησης;

- Καλύτερη **κάλυψη**:
  - Οι σελίδες που είναι γνωστές σε κάθε μηχανή είναι διαφορετικές
- **Διάταξη Πλειοψηφούσας Γνώμης** (consensus ranking)
  - Η διαθεσιμότητα πολλών μηχανών μας δίνει την δυνατότητα να ορίσουμε ένα αθροιστικό (πλειοψηφικό) μέτρο συνάφειας
    - Ενοποίηση αποτελεσμάτων = Πρόβλημα απόφασης ομάδας (group decision problem)
- **Μείωση spam**:
  - Δύσκολα μια spam σελίδα μπορεί να ξεγελάσει όλες τις μηχανές

Μετα-Μηχανές Αναζήτησης

- Examples:
  - Dogpile (http://www.dogpile.com/)
    - over Google, Yahoo!, msn, Ask Jeaves
  - SurfWax (http://www.surfwax.com/)
  - http://www.jux2.com/
  - Metacrawler, SavvySearch,
- Βήματα Λειτουργίας
  - Submit queries to host sites.
  - **Parse resulting HTML pages** to extract search results.
  - **Integrate multiple rankings into a "consensus" ranking**.
  - Present integrated results to user.
- Διαφορές με την Κατανεμημένη Ανάκτηση Πληροφοριών
  - οι υποκείμενες μηχανές δεν παρέχουν term-statistics, άρα μπορούμε να χρησιμοποιήσουμε μόνο απομονωμένες (isolated) τεχνικές ενοποίησης αποτελεσμάτων
  - οι υποκείμενες μηχανές δεν υποστηρίζουν την ίδια ερωτηματική γλώσσα

Ενοποίηση Διατάξεων: **Rank Aggregation (or Meta-Ranking)** (isolated)

Διατύπωση του Προβλήματος

- ***D***: ένα σύνολο αντικειμένων (π.χ. εγγράφων)
- ***S1,…Sk***: ένα σύνολο διατάξεων του D

- Σκοπός: Ενοποίηση των διατάξεων *S1,..Sk σε μία*

The metaphor: **elections**

- Objects → Candidates
- Sources → Electors
- Ordering by a system → Elector's voting ticket
- Fused ordering → Election list

Plurality Ranking (Απλή Πλειοψηφία)

Ο υποψήφιος με τις περισσότερες πρώτες θέσεις είναι ο νικητής…

Έστω 6 πηγές (S1,…,S6) και 4 σελίδες a,b,c,d

S1: <a,c,d,b>
S2: <a,b,c,d>
S3: <b,c,a,b>
S4: <b,a,d,c>
S5: <a,d,c,b>
S6: <c,a,b,d>

a: 3
b: 2
c: 1
d: 0

Τελική κατάταξη: **<a,b,c,d>**

Plurality Ranking (Απλή Πλειοψηφία)
Κάποια προβλήματα

3 συστήματα	<a,c,d,b>		3 συστήματα	<a,c,b>
6 συστήματα	<a,d,c,b>		6 συστήματα	<a,c,b>
3 συστήματα	<b,c,d,a>		3 συστήματα	<b,c,a>
5 συστήματα	<b,d, c, a>	Απόσυρση του d	5 συστήματα	<b,c, a>
2 συστήματα	<c,b,d,a>	(που ήταν τελευταίο) →	2 συστήματα	<c,b,a>
5 συστήματα	<c,d,b,a>		5 συστήματα	<c,b,a>
2 συστήματα	<d,b,c,a>		2 συστήματα	<b,c,a>
4 συστήματα	<d,c,b,a>		4 συστήματα	<c,b,a>

a:9
b:8
c:7
d:6

Τελική διάταξη: **<a,b,c,d>**

a:9
b:10
c:11

Τελική διάταξη: **<c,b,a>**

Αντίστροφη της αρχικής!

Plurality Ranking (Απλή Πλειοψηφία)
Κάποια προβλήματα

3 συστήματα <a,c,d,b>
6 συστήματα <a,d,c,b>
3 συστήματα <b,c,d,a>
5 συστήματα <b,d, c, a>
2 συστήματα <c,b,d,a>
5 συστήματα <c,d,b,a>
2 συστήματα <d,b,c,a>
4 συστήματα <d,c,b,a>

a:9
b:8
c:7
d:6
Τελική διάταξη: **<a,b,c,d>**

Απόσυρση του d
Τελική διάταξη: **<c,b,a>**

Απόσυρση του a
Τελική διάταξη: **<d,c,b>**

Απόσυρση του b
Τελική διάταξη: **<d,c,a>**

Απόσυρση του c
Τελική διάταξη: **<d,b,a>**

Ενοποίηση Διατάξεων κατά **Borda**
[Jean-Charles Borda 1770]

Reinvented (for the context of Meta-Searching) in [Tzitzikas 2001]

The votes of an object o $V(o) = \sum_{i=1..k} r_i(o)$

$r_i(o)$: the position of the object o in the ordering of system S_i

The fused ordering is derived by ordering the objects in *ascending* order wrt to their *votes*

Example:

$S_1 : < o_1, o_2, o_3 >$
$S_2 : < o_1, o_3, o_2 >$
$S_3 : < o_3, o_1, o_2 >$

$V(o_1) = 1+1+2 = 4$
$V(o_2) = 2+3+3 = 8$
$V(o_3) = 3+2+1 = 6$

$M : < o_1, o_3, o_2 >$

If each source S_i returns an ordered *subset* O_i of *Obj*.

$$r_i(o_j) = \begin{cases} \text{position of } o_j \text{ in } O_i, & \text{if } o_j \in O_i \\ F+1 & \text{otherwise} \end{cases} \quad \text{where } F = \max\{|O_1|,...,|O_k|\}$$

Γιατί;

Ενοποίηση Διατάξεων κατά Borda
[Tzitzikas, 2001] **Βαθμός Συμφωνίας**

The *distance* between two orderings *i* and *j:*

$$dist(i, j) = \sum_{o \in O} | r_i(o) - r_j(o) |$$

The *mean distance* of the fused ordering 0

$$Dem = \frac{\sum_{i=1..k} dist(0, i)}{k}$$

The ***level of agreement*** of the fused ordering 0:

- linear transformation $LA = \frac{C - Dem}{C}$ C : max possible mean distance
- inversion transformation $LA = C^{-Dem}$ $C > 1$, e.g. C = 2

- **High** level may drive the user to read only the very first documents since probably they are the more relevant
- **Low** level may drive the user to read more documents

Ενοποίηση Διατάξεων κατά **Condorcet** [1785]

Condorcet: the winner is a candidate that defeats every other candidate in pairwise majority-rule election

S1: <a,b,c>
S2: <b,a,c>
S2: <c,a,b>

a:b 2:1 // ο a νικά τον b δύο φορές (και χάνει μία)
a:c 2:1 // ο a νικά τον c δύο φορές (και χάνει μία)

Condorset ordering: <a,b,c>

Ενοποίηση Διατάξεων κατά **Condorcet** [1785]

S1: <a,b,c>
S2: <b,c,a>
S3: <c,a,b>

a:b 2:1 // άρα ο b δεν μπορεί να είναι ο νικητής
a:c 1:2 // άρα ο a δεν μπορεί να είναι ο νικητής
c:b 1:2 // άρα ο c δεν μπορεί να είναι ο νικητής

Δεν υπάρχει πάντα Condorset νικητής!

Borda vs **Condorcet**

S1: <a,b,c>
S2: <b,a,c>
S2: <c,a,b>

- Condorset
  - a:b 2:1
  - a:c 2:1
  - Condorset ordering: <a,b,c>
- Borda
  - a: 1+2+2 = 5
  - b: 2+1+3 = 6
  - c: 3 + 3 + 1 = 7
  - Borda ordering: <a,b,c>

Borda ≠ Condorcet

Borda (1770)

- Member of French Academy of Sciences
- Noted for work in hydraulics, optics, navigation instrument
- Purpose: Reforming the election procedure of French Academy.
- Criticize plurality method

Condorcet (1785)

- Viewed Borda as an enemy
- Finding best ordering by hypothesis testing
- Switch to propose Condorcet winner

Borda ≠ Condorcet

S1: <a,b,c,d,e>
S2: <b,c,e,d,a>
S3: <e,a,b,c,d>
S4: <a,b,d,e,c>
S5: <b,a,d,e,c>

- Borda
  - a: 1 + 5 + 2 + 1 + 2 = 11
  - b: 2 + 1+ 3 + 2 + 1 = 9
  - c: 3 + 2 + 4 + 5 + 5 =19
  - d: 4 + 4 + 5 + 3 + 3 = 19
  - e: 5+3 +1 + 4 + 4 = 17
  - **Borda winner : b**
- Condorset
  - a:b 3:2
  - a:c 4:1
  - a:d 4:1
  - a:e :3:2
  - **Condorset winner a**

Prurality ≠ Borda ≠ Condorcet

	49 votes	48 votes	3 votes
1*st*	*x*	*y*	*z*
2*nd*	*y*	*z*	*y*
3*rd*	*z*	*x*	*x*

- Prurality winner: x
- Borda winner: y
- Condorcet: z> x

Condorcet and Order

- 3 candidates, 13 voters

	a	*b*	*c*
a	–	8	6
b	5	–	11
c	7	2	–

- <a,b,c> has support 25
  - a>b:8, a>c:6, b>c:11
- <b,c,a> has support 23
  - a<b:5, c>a:7, b>c:11

Ενοποίηση Διατάξεων κατά **Kemeny** (1959)
(Kenemy developed BASIC language)

- Απόσταση μεταξύ δυο διατάξεων = πλήθος των διαφωνιών στη διάταξη ζευγαριών
- Παράδειγμα 1
  - r1 = <a,b,c>
  - r2 = <b, a, c>
  - K(r1, r2) = 1
    - $(a >_{r1} b, a <_{r2} b)$
- Παράδειγμα 2
  - r1 = <a, b, c, d>
  - r2 = <b, d, a, c>
  - K(r1, r2) = 3
    - $(a >_{r1} b, a <_{r2} b)$ $(a >_{r1} d, a <_{r2} d)$ $(c >_{r1} d, c <_{r2} d)$

Ενοποίηση Διατάξεων κατά **Kemeny** (1959)

Kemeny Optimal Aggregation

- Η καλύτερη ενοποιημένη διάταξη είναι εκείνη που απέχει το λιγότερο από όλες τις διατάξεις
- Έστω n διατάξεις: r1, r2, …, rn
- Ενοποιημένη διάταξη $r = \arg\min \sum K(r,ri)$
- Η εύρεση της ενοποιημένης διάταξης είναι ακριβή
  - (πρόβλημα NP-hard)
- Reconciles Borda and Condorcet

Ενοποίηση Διατάξεων: **Επιθυμητές Ιδιότητες**

- Ουδετερότητα (Neutrality)
  - Καμία εναλλακτική δεν πρέπει να ευνοείται
- Pareto Optimality
  - Αν X > Y (σε όλες τις διατάξεις) τότε X>Y (στην τελική)
- Μονοτονία (Monotonicity) // Ranking higher should not hurt a candidate
  - X νικητής (στην τελική), αλλαγή ενός ψηφοδελτίου YZX → YXZ, ο X παραμένει νικητής (στην τελική)
- Ανεξαρτησία από άσχετες εναλλακτικές (Independence from Irrelevant Alternatives)
  - X > Y (στην τελική), αλλαγή ενός ψηφοδελτίου XZY → ZXY, to X>Y παραμένει στην τελική
- Συνέπεια (Consistency)
  - Αν οι ψηφοφόροι διαιρεθούν σε δύο ομάδες και κάθε ομάδα αναδείξει τον ίδιο νικητή, τότε ο τελικός νικητής (αν λάβουμε υπόψη τις ψήφους και των 2 ομάδων) πρέπει να είναι ο ίδιος

Arrow's Impossibility Theorem

Kenneth J. Arrow, *Social Choice and Individual Values* (1951). Won Nobel Prize in 1972

No voting scheme over three or more alternatives can satisfy the following conditions

- Universality (no restriction on individual ordering. All orderings are achievable)
- Monotonicity
- Independence of irrelevant alternatives
- Pareto Optimality
- Non-dictatorship

Arrow's Impossibility Theorem

- Συμπέρασμα: δεν υπάρχει μια απολύτως ικανοποιητική συνάρτηση ενοποίησης διατάξεων