Αξιολόγηση της Ποιότητας των Εξωτερικών Συνδέσμων της Wikipedia

Σοφία Στάμου

Τμήμα Αρχειονομίας και Βιβλιοθηκονομίας
Ιόνιο Πανεπιστήμιο
stamou@ionio.gr

Εισαγωγή

Wikipedia

- Πλούσια, πολυσυλλεκτική πηγή πληροφοριών
- Βέλτιστο παράδειγμα συνεργατικής δράσης στον Ιστό (wiki)
- >10.000.000 άρθρα σε 270 γλώσσες
- >13.000.000 συγγραφείς
- 1.750 διαχειριστές
- >78.000.000 επισκέπτες / μήνα

Ποια η ποιότητα και η εγκυρότητά της;

Σχετική έρευνα

Ποιοτικά χαρακτηριστικά περιεχομένου:

- Έκταση άρθρων & κατανομή διακριτών λέξεων [Voss 05]
- Συχνότητα επικαιροποίησης άρθρων [Wilkinson & Huberman 07]
- Αφοσίωση [Riehle, 2005], διαφωνίες επιμελητών [Vuong et al. 08]
- Εξερχόμενοι σύνδεσμοι σε «αξιόπιστες» πηγές [Nielsen 07]
- Αντικειμενικότητα περιεχομένου [Kirtsis et al. 11]

Ποιοτικά χαρακτηριστικά δομής:

- Συνεκτικότητα γράφου [Koolen & Kamps 09]
- Κύκλοι πλοήγησης [Kirtsis et al. 10]
- Πληρότητα και ποιότητα κατηγοριών [Sakai et al. 09]

Διασφάλιση ποιότητας

Προκλήσεις

- Κριτήρια, μετρικές αξιολόγησης
- Υποστήριξη συγγραφής και δόμησης περιεχομένου
- Συντονισμένος περιοδικός έλεγχος (αυτοματοποιημένος)
- Επικαιροποίηση περιεχομένου και δομής
- Προβολή «ποιοτικού» περιεχομένου στους χρήστες

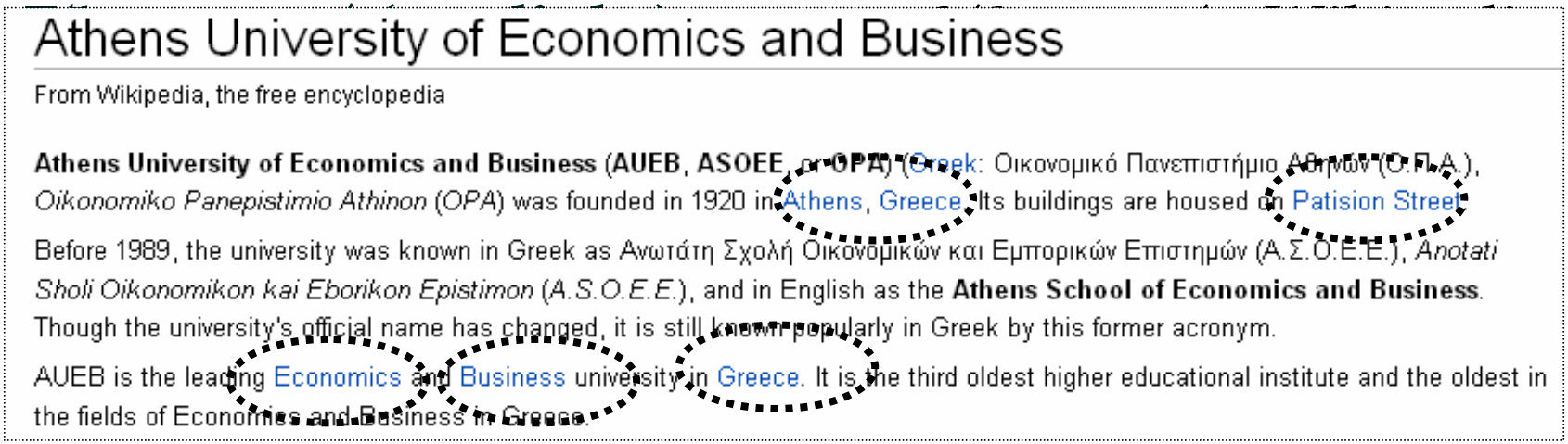
Περιεχόμενα Ομιλίας

- Συμβολή εξωτερικών συνδέσμων στην ποιότητα του περιεχομένου της Wikipedia
  - Τύποι και σημασιολογία υπερσυνδέσμων
  - Επικάλυψη περιεχομένου άρθρου- εξωτερικών πόρων
- Κανόνες διασύνδεσης άρθρων με εξωτερικές πηγές
  - Επικαιροποίηση υπερσυνδέσμων και διασφάλιση εγκυρότητας περιεχομένου

Υπερσύνδεσμοι στη Wikipedia

Εσωτερικοί (in-links): προς άρθρα της Wikipedia

- Συμπληρωματικότητα άρθρων
- Αυτονομία εγκυκλοπαίδειας



Athens University of Economics and Business

From Wikipedia, the free encyclopedia

Athens University of Economics and Business (**AUEB**, **ASOEE**, or **OPA**) (Greek: Οικονομικό Πανεπιστήμιο Αθηνών (Ο.Π.Α.), *Oikonomiko Panepistimio Athinon* (*OPA*) was founded in 1920 in Athens, Greece. Its buildings are housed on Patision Street.

Before 1989, the university was known in Greek as Ανωτάτη Σχολή Οικονομικών και Εμπορικών Επιστημών (Α.Σ.Ο.Ε.Ε.), *Anotati Sholi Oikonomikon kai Eborikon Epistimon* (*A.S.O.E.E.*), and in English as the **Athens School of Economics and Business**. Though the university's official name has changed, it is still known popularly in Greek by this former acronym.

AUEB is the leading Economics and Business university in Greece. It is the third oldest higher educational institute and the oldest in the fields of Economics and Business in Greece.

Υπερσύνδεσμοι στη Wikipedia

Εξωτερικοί (out-links): προς σελίδες εκτός Wikipedia

- Επιπλέον πληροφορίες σχετικού περιεχομένου

Αναφορών (references): προς βιβλιογραφικές πηγές

- Εγκυρότητα περιεχομένου

Κίνητρο Μελέτης

Η ποιότητα του περιεχομένου της Wikipedia εξαρτάται **και** από την ποιότητα των υπερσυνδέσμων των άρθρων της

Δείκτες Ποιότητας

- Αποφυγή κύκλων (in-links)
- Αξιοπιστία περιεχομένου (references)
- Συμπληρωματική και επικαιροποιημένη πληροφορία (out-links)

Πρόκληση: αυτοματοποιημένη αξιολόγηση της ποιότητας των εξωτερικών υπερσυνδέσμων

Τι εξετάζουμε....

- Κατανομή υπερσυνδέσμων
  - Συσχέτιση έκτασης άρθρου και πλήθους υπερσυνδέσμων
  - Δημοφιλή domains
- Παρακμή υπερσυνδέσμων
  - Αναλογία dead links στους υπερσυνδέσμους των άρθρων
  - Soft-404 server errors (redirects)
- Συνεισφορά υπερσυνδέσμων
  - Συμπληρωματικότητα περιεχομένου
  - «νέα» πληροφορία

Κατανομή υπερσυνδέσμων

- Εξόρυξη out-links από dump files
- Απαλοιφή space, bullet χαρακτήρων
- Συσταδοποίηση out-links βάσει σύνταξης

[URLs] + hyperlinks

```
"The [[RFC]]-mandated [http://example.com/ example.com website]"
"The RFC-mandated example.com website".
```

Hyperlinked URL names

```
[http://example.com/]
[2]
```

[URLs] + επισημείωση hyperlinks

```
[http://example.com/ The RFC-mandated example.com website]
The RFC-mandated example.com website
```

Κατανομή υπερσυνδέσμων (2)

- Εξόρυξη URLs από κάθε συστάδα και αναγνώριση domain names
- Επεξεργασία XML files και εξόρυξη περιεχομένου
- Υπολογισμός πλήθους χαρακτήρων κάθε άρθρου
- Αναπαράσταση κειμένου και out-links

  <μήκος άρθρου, πλήθος out-links>

Παρακμή υπερσυνδέσμων

Dead pages (hard 404-errors)

- url invalid or can't be parsed
- Local DNS server can't resolve IP of url.HOST
- No response from url.HOST within T(=10) secs
- url.HOST returns error HTTP code to the request for url

```
Function bool isDeadPage(u)
in: URL u
1:  string T_u, T_r, int K_u, K_r, bool error
2:  fetch(u, w_u, T_u, K_u, error)
3:  if (error) then    // A hard-404
4:    return true

Function atomicFetch(w, T, v, redirect, error)
in: URL w
out: string T, URL v, bool redirect, bool error
1:  parse(w, error)
2:  if (error) then    // parse URL failed
3:    return
4:  IPAddress address
5:  getIPAddress(w.HOST, address, error)
6:  if (error) then    // resolution of host's IP address failed
7:    return
8:  HTTPRetCode code
9:  httpGet(address, T, v, code, timeout = 10sec, error)
10: if (error) then    // http get timed out
11:   return
12: if (code in { 403, 404, 410, 5xx} ) then
13:   error = true; return
```

Παρακμή υπερσυνδέσμων (2)

Decayed pages (soft 404-errors)

- url redirects to its host's page and a non-existing randomly generated url in the same host also redirects to the host's page
- url redirects to a page, which in turn redirects (loop of redirects)
- url and a non-existing randomly generated url redirect to pages of identical content

```
5:  URL r = u.PARENT+ 25 random characters
6:  fetch(r, w_r, T_r, K_r, error)
7:  if (error) then                          → host returns hard 404
8:    return false
9:  if (u is the root of u.HOST) then
10:   return false                           → root can't be soft 404
11: if(K_u ≠ K_r) then                       → different redirects
12:   return false
13: if (w_u = w_r) then                      → same redirects
14:   return true
15: if (shingle(T_u) = shingle(T_r)) then
16:   return true
17: return false    // not a soft-404 page
```

Συνεισφορά υπερσυνδέσμων

- Προσκομιδή του περιεχομένου των εξωτερικών συνδέσμων κάθε άρθρου και επεξεργασία
- Επεξεργασία περιεχομένου άρθρων
- Εφαρμογή shingling [Broder, 97] , υπολογισμός Containment στο κείμενο άρθρων και εξωτερικών συνδέσμων

$$Containment\left(a_i, d_i\right) = \frac{\left|S\left(a_i\right) \cap S\left(d_i\right)\right|}{\left|S\left(a_i\right)\right|}$$

Shingles άρθρου

Shingles out-linked σελίδας άρθρου

Συνεισφορά υπερσυνδέσμων

Avg. Containment άρθρου με όλες τα out-linked docs

$$avg.Containment(a_i) = \frac{1}{|D|}\sum_{d_i \in D}^{|D|} Containment(a_i, d_j), \ldots, (a_i, d_n)$$

Πλήθος out-linked docs

Uniqueness περιεχομένου άρθρου συγκριτικά με τα out-linked docs

$$Uniqueness(a_i) = 1 - avg.Containment(a_i)$$

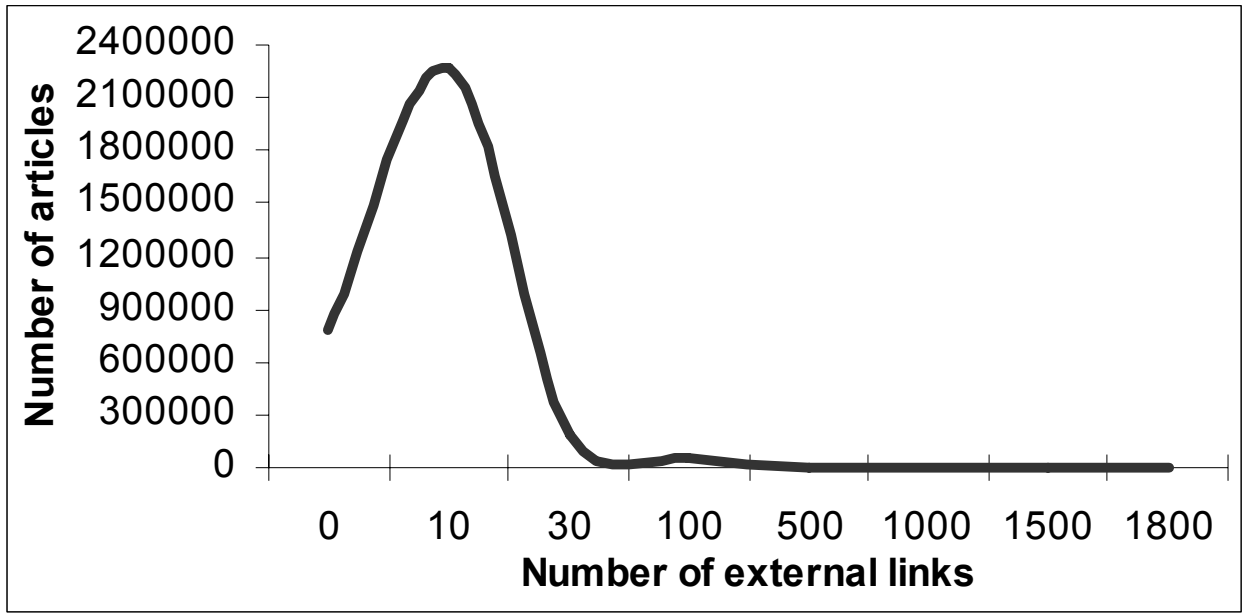
Πειραματική Αξιολόγηση
Κατανομή out-links

Collection ID	October 2009 Wikipedia dump
Number of articles	3,290,179
Number of external links	13,355,687
External links/ article	4.06
Articles without external links	786,857

23.91% των άρθρων χωρίς out-links

Προσκομιδή των 2,503,322 άρθρων με out-links, επεξεργασία περιεχομένου, υπολογισμός μεγέθους (πλήθος χαρακτήρων)

Κατανομή out-links

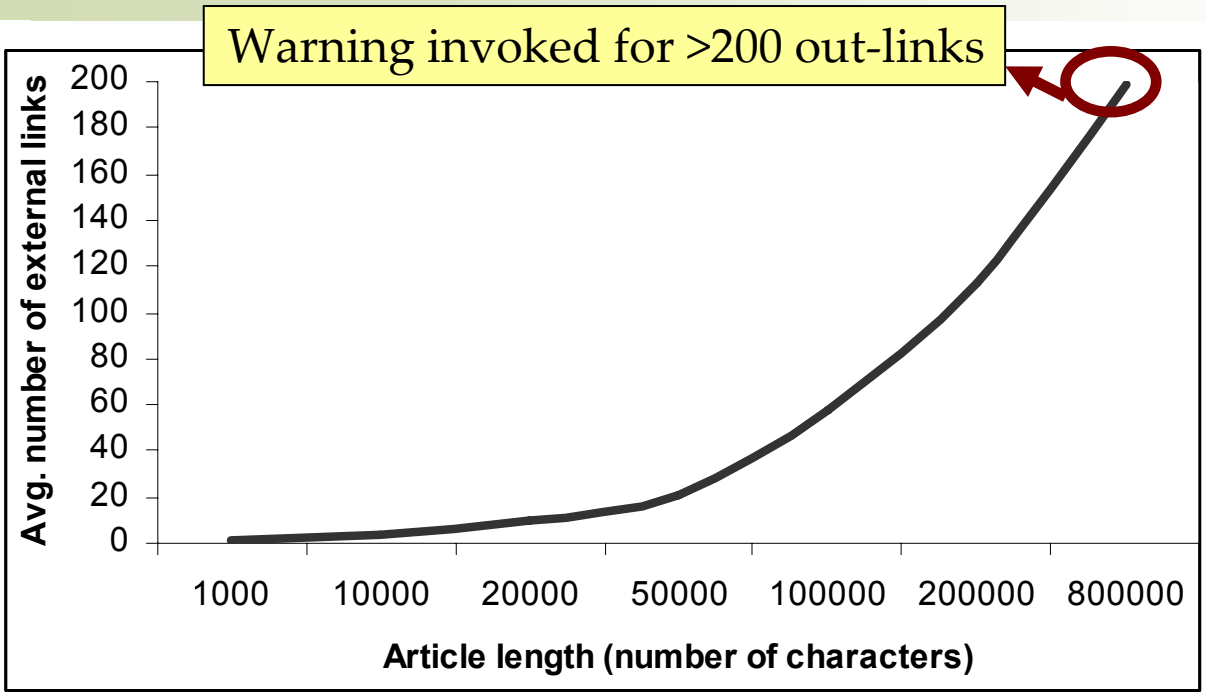


- 7.30% άρθρα με > 100 out-links (ημιτελής πληροφορία;)
- 23.91% άρθρα χωρίς out–links (ολοκληρωμένη πληροφορία;)
- 68.79% άρθρα 1-10 out-links (συμπληρωματικότητα;)

⬇

Τήρηση οδηγιών παράθεσης external links από συγγραφείς

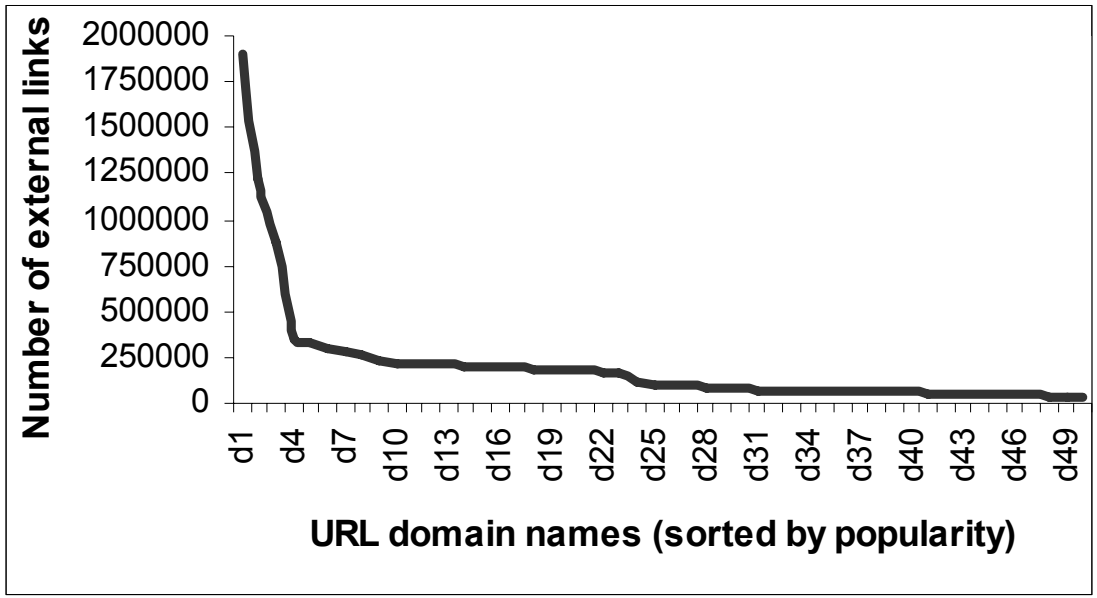
Μέγεθος άρθρων και out-links



- Συσχέτιση μεγέθους κειμένου και πλήθους out-links
- 70.6% των άρθρων έχουν 10K χαρακτήρες και 3 έως 8 out-links

Περισσότερη πληροφορία → περισσότερες πηγές πληροφόρησης
Συμπληρωματικότητα (+) / Αυτοτέλεια (-)

Κατανομή out-link domains



- Το πιο δημοφιλές domain φιλοξενεί 14.29% των wiki out-links
- 44.87% των out-links σε 10 διαφορετικά domain names

Υψηλή θεματική συσχέτιση και εσωτερική συνδεσιμότητα μεταξύ άρθρων [Koolen & Kamps,09] → αύξηση δημοτικότητας domains

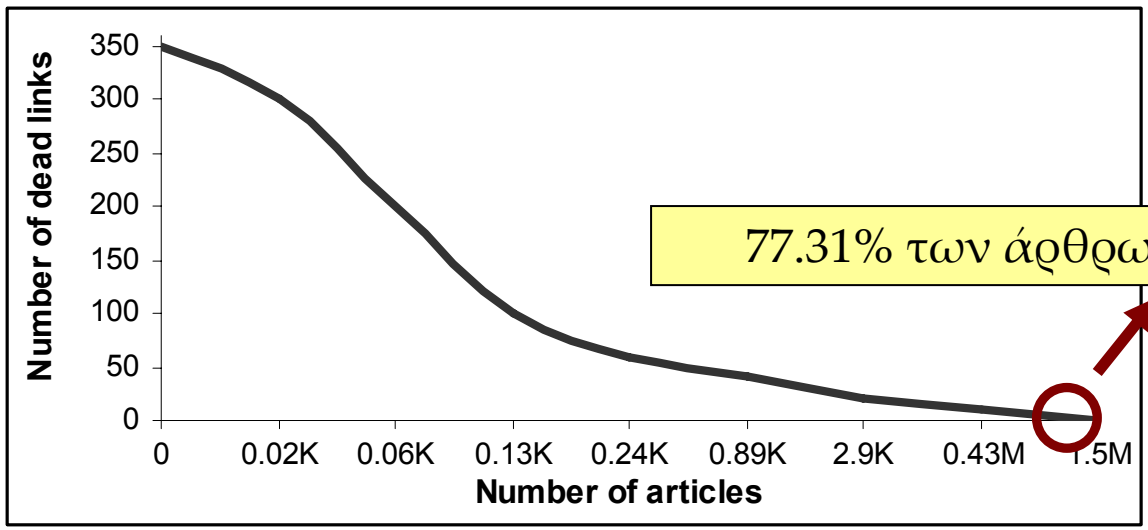
Πειραματική Αξιολόγηση
Παρακμή out-links

Δεδομένα και διαδικασία αξιολόγησης

Articles examined	1,939,445
Out-links evaluated	4,575,154
Page fetching time out	10 secs
Max # of redirects	5
Characters appended to host's url	25 lower case Latin

Αλγόριθμος υπολογισμού hard- (dead) & soft-404 (redirect) page errors

Κατανομή dead out-links



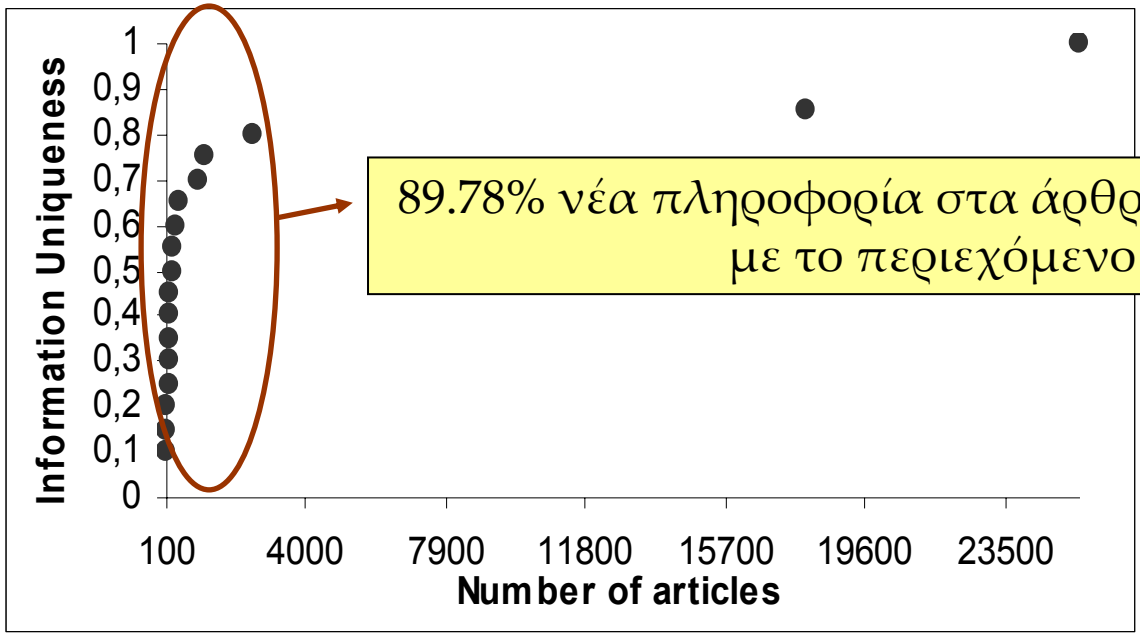
77.31% των άρθρων χωρίς decayed out-link

- 18.34% των links που εξετάσαμε είναι dead (hard or soft-404 code)

Εγκυρότητα πληροφορίας - Ορθή επιλογή υπερσυνδέσμων

Συνεισφορά out-links

Σε τυχαίο δείγμα 100,000 άρθρων



Συμπληρωματικότητα περιεχομένου

Συμπεράσματα μελέτης

- Τήρηση κανόνων συνδεσιμότητας άρθρων – ιστοσελίδων βάσει ποσοτικών και ποιοτικών κριτηρίων
- Ανάγκη αυτοματοποιημένου ελέγχου και επικαιροποίησης out-links
- Wikipedia: συλλεκτική πηγή πληροφοριών – διαφορετικό περιεχόμενο

Δημοσιευμένη σχετική έρευνα

Wikipedia Quality Assessment

N. Kirtsis, P. Tzekou, **S. Stamou**, N. Zotos. *Information Uniqueness in Wikipedia.* [**WebIST 2010**]

N. Kirtsis, J. Besharat, P. Tzekou, **S. Stamou**. *Identifying Polarized Wikipedia Articles.* [**Ontology-Driven Web Mining 2011**]

P. Tzekou, **S. Stamou**, N. Kirtsis, N. Zotos. *Quality Assessment of Wikipedia External Links.* [**WebIST 2011**]

Έρευνα σε εξέλιξη

- **Scientometrics**
  - Quality Assessment Metrics for Research Conferences
  - Interpreting Self-Citations in Scientific Publications
  - Mining Citation Opinion Terms [JDIM10]
  - Impact of Funded Research Works [ICDM09]
- **Web Search**
  - Query Rewrites for Task-Oriented Web Searches
  - Abandoned Web Searches [ECIR10, SIGIR09]

Ερωτήσεις…

Ευχαριστώ!