Βάσεις Διαδικτύου

Θέματα

Εισαγωγή στην XML

Ευρετήρια για την Ανάκτηση Κειμένων

Ο αλγόριθμος HITS

Τι είναι η XML

Mark-up Γλώσσες (Γλώσσες Σημειοθέτησης)

HTML ετικέτες (tags) για την αναπαράσταση της δομής των εγγράφων

XML (Extensible Markup Language) δε διαθέτει μια συγκεκριμένη συλλογή ετικετών με σταθερή και καθορισμένη σημασία

Αντίθετα, ο χρήστης μπορεί να ορίσει δικές του ετικέτες – που συνήθως αφορούν τη σημασία του περιεχομένου

Παράδειγμα XML

```
<BOOK>
        <AUTHOR>
                <FIRSTNAME>Richard</FIRSNAME>
                <LASTNAME>Feymann</LASTNAME>
        </AUTHOR>
</BOOK>
```

Απλά και εμφωλευμένα στοιχεία (elements/sub-elements)

Ετικέτα αρχής (<elm>) και ετικέτα τέλους (</elm>)

Παράδειγμα XML

```
<BOOK genre="Science" format="Hardcover">
        <AUTHOR>
                <FIRSTNAME>Richard</FIRSNAME>
                <LASTNAME>Feymann</LASTNAME>
        </AUTHOR>
</BOOK>
```

Γνωρίσματα (Attributes) περιγραφικές πληροφορίες για τα στοιχεία

att = "value"


```
<?XML version="1.0" encoding="UTF-8" standalone="yes"?>
<!DOCTYPE BOOKLIST SYSTEM "emp.dtd">
<BOOKLIST>
<BOOK genre="Science" format="Hardcover">
        <AUTHOR>
                <FIRSTNAME>Richard</FIRSNAME>
                <LASTNAME>Feymann</LASTNAME>
        </AUTHOR>
        <TITLE>The Character of Physical Law</TITLE>
        <PUBLISHED>1980</PUBLISHED>
</BOOK>
<BOOK genre="Fiction">
        <AUTHOR>
                <FIRSTNAME>R. K</FIRSNAME>
                <LASTNAME>Narayan</LASTNAME>
        </AUTHOR>
        <TITLE>The Character of Physical Law</TITLE>
        <PUBLISHED>1981</PUBLISHED>
</BOOK>
```

Παράδειγμα XML

```
<BOOK genre="Fiction">
        <AUTHOR>
                <FIRSTNAME>R. K</FIRSNAME>
                <LASTNAME>Narayan</LASTNAME>
        </AUTHOR>
        <TITLE>The English Teacher</TITLE>
        <PUBLISHED>1980</PUBLISHED>
</BOOK>
</BOOKLIST>
```

Ορθά Διαμορφωμένο (well-formed)

- Ξεκινά με δηλωτικό
- Υπάρχει στοιχείο ρίζα
- Κατάλληλα εμφωλευμένα στοιχεία

XML DTD

Μια δήλωση DTD είναι ένα σύνολο από κανόνες που επιτρέπουν στο χρήστη να ορίζει το δικό του σύνολο στοιχείων και γνωρισμάτων

Έγκυρο έγγραφο όταν συνοδεύεται από μια αντίστοιχη DTD και το έγγραφο είναι δομημένο σύμφωνα με τους κανόνες που ορίζει η DTD

Παράδειγμα DTD

```
<!DOCTYPE BOOKLIST [
<!ELEMENT BOOKLIST (BOOK)*>
        <!ELEMENT BOOK (AUTHOR, TITLE, PUBLIHSED?)>
                <!ELEMENT AUTHOR (FISRTNAME, LASTNAME)>
                        <!ELEMENT FIRSTNAME(#PCDATA)>
                        <!ELEMENT LASTNAME (#PCDATA)>
                <!ELEMENT TITLE (#PCDATA)>
                <!ELEMENT PUBLISHED (#PCDATA)>
<!ATTLIST BOOK genre (Science | Fiction) #REQUIRED>
<!ATTLIST BOOK format (Paperback | Hardcover) "Paperback">
]>
```


Παράδειγμα DTD

```
<!DOCTYPE BOOKLIST [                                  Στοιχείο ρίζα
<!ELEMENT BOOKLIST (BOOK)*>
        <!ELEMENT BOOK (AUTHOR,TITLE, PUBLIHSED?)>
                <!ELEMENT AUTHOR (FISRTNAME, LASTNAME)>
                        <!ELEMENT FIRSTNAME(#PCDATA)>
                        <!ELEMENT LASTNAME (#PCDATA)>
                <!ELEMENT TITLE (#PCDATA)>
                <!ELEMENT PUBLISHED (#PCDATA)>
<!ATTLIST BOOK genre (Science | Fiction) #REQUIRED>
<!ATTLIST BOOK format (Paperback | Hardcover) "Paperback">

]>
```

Γενικό σχήμα <!DOCTYPE *name* [

DTDDeclaration

] >

Παράδειγμα DTD

```
<!DOCTYPE BOOKLIST [                                  subelements
<!ELEMENT BOOKLIST (BOOK)*>
        <!ELEMENT BOOK (AUTHOR, TITLE, PUBLIHSED?)>
                <!ELEMENT AUTHOR (FISRTNAME, LASTNAME)>
                        <!ELEMENT FIRSTNAME(#PCDATA)>
                        <!ELEMENT LASTNAME (#PCDATA)>
                <!ELEMENT TITLE (#PCDATA)>
                <!ELEMENT PUBLISHED (#PCDATA)>
<!ATTLIST BOOK genre (Science | Fiction) #REQUIRED>
<!ATTLIST BOOK format (Paperback|Hardcover) "Paperback">
]>
```

* 0 ή περισσότερα

+ 1 ή περισσότερα

? Προαιρετική εμφάνιση,

Παράδειγμα DTD

```
<!DOCTYPE BOOKLIST [                                  subelements
<!ELEMENT BOOKLIST (BOOK)*>
        <!ELEMENT BOOK (AUTHOR, TITLE, PUBLIHSED?)>
                <!ELEMENT AUTHOR (FISRTNAME, LASTNAME)>
                        <!ELEMENT FIRSTNAME(#PCDATA)>
                        <!ELEMENT LASTNAME (#PCDATA)>
                <!ELEMENT TITLE (#PCDATA)>
                <!ELEMENT PUBLISHED (#PCDATA)>
<!ATTLIST BOOK genre (Science | Fiction) #REQUIRED>
<!ATTLIST BOOK format (Paperback|Hardcover) "Paperback">
]>
```

* 0 ή περισσότερα

+ 1 ή περισσότερα

? Προαιρετική εμφάνιση,

Παράδειγμα DTD

```
<!DOCTYPE BOOKLIST [
<!ELEMENT BOOKLIST (BOOK)*>
        <!ELEMENT BOOK (AUTHOR, TITLE, PUBLIHSED?)>
                <!ELEMENT AUTHOR (FISRTNAME, LASTNAME)>
                        <!ELEMENT FIRSTNAME(#PCDATA)>
                        <!ELEMENT LASTNAME (#PCDATA)>
                <!ELEMENT TITLE (#PCDATA)>
                <!ELEMENT PUBLISHED (#PCDATA)>
<!ATTLIST BOOK genre (Science | Fiction) #REQUIRED>
<!ATTLIST BOOK format (Paperback|Hardcover) "Paperback">
]>
```

#PCDATA δηλώνει στοιχεία με μορφή χαρακτήρων

Παράδειγμα DTD

Γενικά

<!ELEMENT (contenttype) >

Όπου contenttype

- Άλλα στοιχεία
- #PCDATA
- EMPTY
- Κανονική έκφραση

  exp1, exp2, exp3, …

  exp*

  exp?

  exp+

  exp1 | exp2

Παράδειγμα DTD

```
<!DOCTYPE BOOKLIST [
<!ELEMENT BOOKLIST (BOOK)*>
        <!ELEMENT BOOK (AUTHOR, TITLE, PUBLIHSED?)>
                <!ELEMENT AUTHOR (FISRTNAME, LASTNAME)>
                        <!ELEMENT FIRSTNAME(#PCDATA)>
                        <!ELEMENT LASTNAME (#PCDATA)>
                <!ELEMENT TITLE (#PCDATA)>
                <!ELEMENT PUBLISHED (#PCDATA)>
<!ATTLIST BOOK genre (Science | Fiction) #REQUIRED>
<!ATTLIST BOOK format (Paperback|Hardcover) "Paperback">
]>
```

<!ATTLIST elementName (attName attType default)+>

attType: τύποι απαρίθμησης ή τύποι συμβολοσειράς

<!ATTLIST BOOK edition CDATA "1">

XML

XQuery: Γλώσσα ερωτήσεων για XML δεδομένα

Τεχνικές

- Για την αποθήκευση δεδομένων σε σχεσιακές βάσεις δεδομένων
- Ειδικές (native) βάσεις δεδομένων για αποθήκευση εγγράφων XML

http://www.w3.org/XML/

http://www.w3.org/XML/Query

Θέματα

Εισαγωγή στην XML

→ Ευρετήρια για την Ανάκτηση Κειμένων

Ο αλγόριθμος HITS

Ευρετηριοποίηση για την Ανάκτηση Κειμένων

Βάση κειμένων: συλλογή από έγγραφα

Αναζήτηση με μια λέξη κλειδί (keyword queries)

Αίτημα Boole

$(t_{11} \vee t_{12} \vee \ldots \vee t_{i1}) \wedge (t_{21} \vee t_{22} \vee \ldots \vee t_{2i2}) \wedge \ldots . (t_{j1} \vee t_{j2} \vee \ldots \vee t_{jij})$

Αίτημα Διαβάθμισης (Ranking)

Ευρετηριοποίηση για την Ανάκτηση Κειμένων

Παράδειγμα

Rid	Λέξεις-Κλειδιά
1	agent James Bond
2	agent mobile computer
3	James Madison movie
4	James Bond movie

Παράδειγμα ερωτήσεων

Ανεστραμμένο Αρχείο

Μια ταξινομημένη λίστα (ανεστραμμένη λίστα) για κάθε όρο

Ευρετήριο Λεξιλογίου:

Για τον ταχύτερο εντοπισμό της λίστας για κάθε όρο: Το σύνολο των όρων μπορεί να οργανωθεί με τη χρήση μιας δομής ευρετηρίου (π.χ. B+-δέντρο)

Παράδειγμα

Ένας όρος, σύζευξη, διάζευξη

Αρχείο Υπογραφών

Υπογραφή εγγράφου (File Signature) Μια εγγραφή ευρετηρίου για κάθε έγγραφο στη βάση δεδομένων

Σταθερό μήκος bits – εύρος υπογραφής

Υ1 ταιριάζει με Υ2, Υ1 τουλάχιστον τα 1 που έχει και η Υ2

Εσφαλμένη διάγνωση (false positive)

Παράδειγμα

Ένας όρος, σύζευξη, διάζευξη

Αρχείο Υπογραφών

Αρχείο υπογραφών με κατακόρυφο διαμερισμό σε μονοψήφιες στήλες:

Διαμερίζουμε ένα αρχείο υπογραφών σε ένα σύνολο κατακόρυφων δυαδικών στηλών

Για κ άσσους ανάκτηση κ-στηλών

Θέματα

Εισαγωγή στην XML

Ευρετήρια για την Ανάκτηση Κειμένων

→ Αναζητήσεις λέξεων κλειδιών στο διαδίκτυο: Ο αλγόριθμος HITS

Ο Αλγόριθμος HITS

Δύο τύποι σελίδων

Αυθεντική

Μια σελίδα που είναι αυθεντία σε ένα θέμα και αναγνωρίζεται ως τέτοια από άλλες σελίδες (δηλαδή, υπάρχουν πολλοί σύνδεσμοι σε αυτήν)

Κομβικοί

Μια σελίδα που αναφέρεται σε μια αυθεντική σελίδα

Ο Αλγόριθμος HITS

Το web ως ένας κατευθυνόμενος γράφος

Κόμβοι: ιστοσελίδες

Ακμή από *A* στον *B*: η ιστοσελίδα *A* έχει έναν υπερ-σύνδεσμο στην ιστοσελίδα *B*

Ο αλγόριθμος σε 2 φάσεις:

Φάση Ι: (δειγματοληπτικό στάδιο) ένα σύνολο σελίδων που αποτελεί το βασικό σύνολο

Φάση ΙΙ: (επαναληπτικό στάδιο) επεξεργασία του βασικού συνόλου για τον εντοπισμό καλών αυθεντικών και κομβικών ιστοσελίδων

Ο Αλγόριθμος HITS

Φάση Ι: Υπολογισμός βασικού συνόλου

1. Υπολογισμός αρχικού συνόλου: σύνολο-ρίζα

Κλασικοί μέθοδοι: πχ ανάκτηση όλων των σελίδων που περιέχουν τις λέξεις κλειδιά

(περιμένουμε ότι θα περιέχει (τουλάχιστον) αναφορές προς σχετικές σελίδες)

2. Σελίδες-σύνδεσμοι: σελίδα που είτε συμπεριλαμβάνει σύνδεσμο που να αναφέρεται στο σύνολο ρίζα είτε το σύνολο ρίζα περιέχει σύνδεσμο που αναφέρεται σε αυτήν

Βασικό Σύνολο: διεύρυνση του συνόλου-ρίζα ώστε να περιλαμβάνει και τις σελίδες συνδέσμους – *Βασικές ιστοσελίδες*

Ο Αλγόριθμος HITS

Φάση ΙΙ: Ποιες βασικές ιστοσελίδες είναι κόμβοι και αυθεντίες

Κάθε βασική σελίδα p δύο τιμές:

h_p - Συντελεστής Κομβικού Ρόλου (πολλούς δείκτες σε αυθεντικές)

a_p - Συντελεστής Αυθεντικότητας (πολλοί δείκτες από κομβικές σε αυτήν)

Αρχικοποίηση, $\forall$ p, $h_p = 1$ και $a_p = 1$

Επαναληπτικά, αυξάνεται

$$a_p = \sum_{\text{Βασικές σελίδες q που δείχνουν στην p}} h_q$$

$$h_p = \sum_{\text{Βασικές σελίδες q στις οποίες δείχνει η p}} a_q$$

Ο Αλγόριθμος HITS

Έστω το βασικό σύνολο σελίδων {1, 2, …, n}

Πίνακας Γειτνίασης (adjacency matrix) B: n x n

B[i, j] = 1 αν η σελίδα i περιέχει σύνδεσμο που δείχνει στη σελίδα j

Έστω $h = \langle h_1, h_2, \ldots, h_n \rangle$ το διάνυσμα συντελεστών κομβικών ρόλων

και $a = \langle a_1, a_2, \ldots, a_n \rangle$ το διάνυσμα συντελεστών αυθεντικότητας

Ο Αλγόριθμος HITS

Οι κανόνες ενημέρωσης

$h = B\,a$ $\qquad$ $a = B^T h$

1η επανάληψη

$h = B\,B^T h = (B\,B^T)\,h$ $\qquad$ $a = B^T B\,a = (B^T B)\,a$

2η επανάληψη

$h = (B\,B^T)^2\,h$ $\qquad$ $a = (B^T B)^2\,a$

Σύγκλιση στα ιδιοδιανύσματα του BB^T και B^TB αν κανονικοποιηθούν αρχικά οι συντελεστές