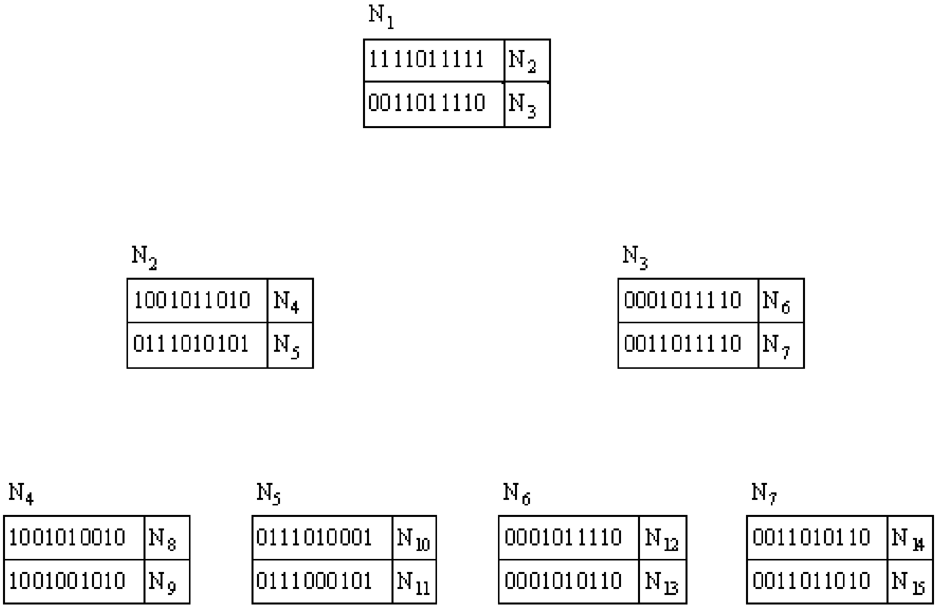
Brandenburgische
Technische Universität
Cottbus - Senftenberg

Database and Information Systems Group
Module: Information Retrieval

Task Sheet 4
Signature Files, Extended Boolean Retrieval, Fuzzy Set Model

1. Explain the idea of using *signatures* in an IR system! What are the advantages or disadvantages of this technique?
2. Describe the method of *bit slice organization* for the management of signatures. What advantages does this have in comparison to *sequential signature files*?
3. What is *superimposed coding* and how can it lead to *false drops* when using signatures?
4. Given the *S-tree* below, explain the search for the dataset `0001011110`!



5. Explain the procedure of adding a new dataset into an S-tree!
6. Explain the *coordination level match retrieval model* and give advantages and disadvantages in comparison to the Boolean model.
7. Name the main advantages of the *p-norm retrieval model* over the traditional Boolean model.
8. What is the role of the choice of the *p* parameter in the *p*-norm model and how does the parameter affect the query results?
9. How does the *fuzzy set model* differ from the Boolean retrieval model? How do disjunction, conjunction and negation work in the fuzzy set model?

10. You are given an IR system that uses the fuzzy set model. The document collection contains four documents D1-D4, with the content being:

    D1: Cottbus is the only city in Lusatia with a basketball team.

    D2: The city of Cottbus ist crazy for Basketball.

    D3: Basketball is not important in Senftenberg.

    D4: Senftenberg and Cottbus are in Lusatia.

    Now assume that after a conversion into lowercase, stop word elimination and stemming we have a vocabulary with the terms *{cottbus, city, lusatia, basketball, team, senftenberg}*.

    Determine first the *term×term correlation matrix* for the vocabulary. Then use it to calculate the membership values $\mu_{i,j}$ for all documents D_j to all terms t_i.

11. Using the membership values calculated in task 10, determine the fuzzy sets for the following queries:

    - Q_1: Senftenberg $\wedge$ Cottbus $\wedge$ Lusatia
    - Q_2: (Senftenberg $\vee$ Cottbus) $\wedge$ $\neg$Basketball
    - Q_3: Senftenberg $\vee$ Lusatia $\vee$ city