Searching for Similar Documents Between Mobile Devices

Kristóf Csorba and István Vajk

This paper proposes a system helping mobile device users to search for documents which are similar to the ones already stored on the user's device locally. The goal is to create a background process monitoring available documents in a peer-to-peer environment and notifying the user if a new document of possible interest gets retrievable. As the system is designed for mobile device environment, it has to satisfy strict conditions according processor and memory consumption. The program running in background may not cause significant performance degradation, otherwise the user interface might get slower which decreases usability. This is accomplished by employing very fast and simple algorithms for topic representation and comparison. Beside the resource consumption, the communication traffic created by the mobile device has to be kept low as well, because traffic usually costs a fee. Users would probably not like to pay huge bills just because background processes are transmitting document topic representations between the mobile devices. The compact topic representation is achieved using topic specific keyword lists and mask vectors indicating the presence or absence of a given keyword in the document. These topic representations are then compared using a simple similarity measure based on the number of common keywords. The key idea behind the compact topic representation is an iterative selection of the best matching keyword list, that is, the keyword list having the most common keywords with the document. After this keyword list is selected, all the contained keywords get a binary flag indicating their presence or absence in the document. As the keyword lists are globally available, topic representations contain only a keyword list identifier number and a single bit for all keywords in it in a form of a binary mask vector. If a keyword list has not more than 128 keywords and the keyword list identifier is 16 bit, the topic representation is 16+128 bit = 18 byte for every document. Using the keyword lists, these mask vectors can be mapped into a global space of all possible keywords and compared using simple scalar product resulting the number of common keywords between the observed documents. As topic specific keywords are relative rare, there are many documents not containing any of the keywords which makes the topic unidentifiable. To overcome this limitation, a cascaded selection procedure is employed using multiple levels of document selection: after documents similar to the local ones are retrieved, the keyword lists are extended with additional keywords (usually with lower quality). This allows the retrieval of further documents, although with lower precision. If the user is allowed to choose the allowed amount of additional keywords, a trade-off between precision and recall may be defined by the user. This paper presents the techniques described above in detail, together with experimental results investigating multiple aspects of the proposed systems performance.