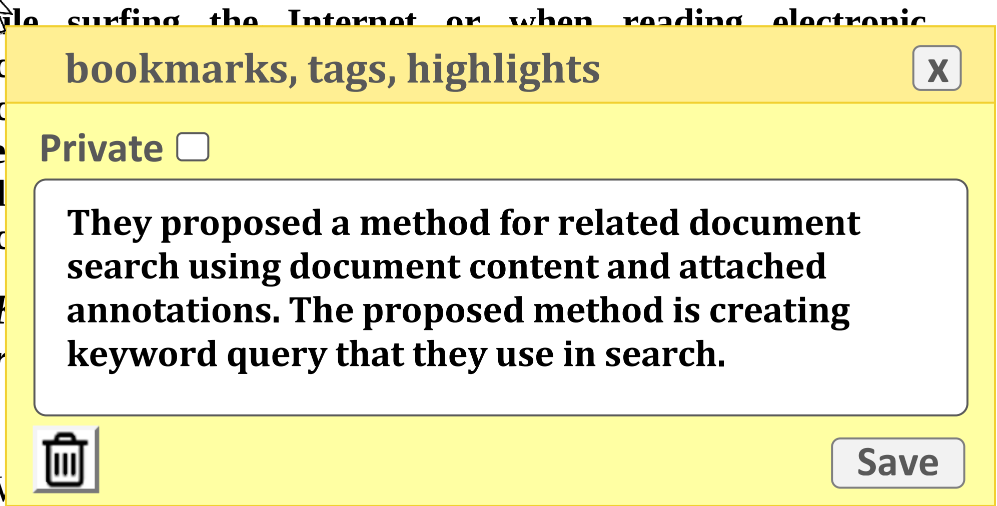
Related Document Search Using User Created Annotations

Jakub Ševcech
supervised by prof. Maria Bielikova
Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova, 842 16 Bratislava, Slovakia
E-mail: sevo.jakub@gmail.com

Abstract—**We often use various services for creating bookmarks, tags, highlights and other types of annotations while surfing the Internet or when reading electronic doc
doc
use
ind
doc**

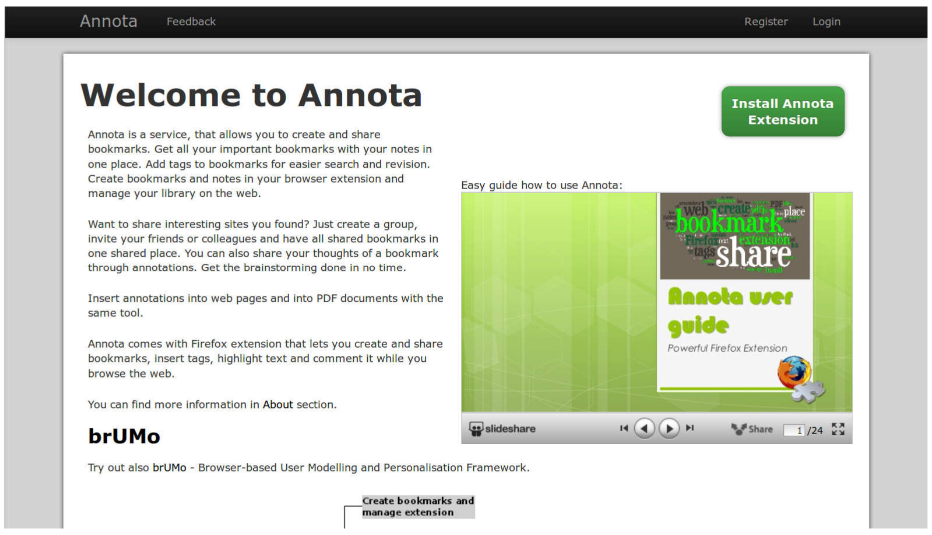
K
spr



They proposed a method for related document search using document content and attached annotations. The proposed method is creating keyword query that they use in search.

V

to insert various types of annotations into web pages and PDF documents displayed in browser. We have developed an extension for Firefox, which allows users highlight text bookmark pages and organizing them by tags, attach comments to text selections and attach notes to documents.



Welcome to Annota

Annota is a service, that allows you to create and share bookmarks. Get all your important bookmarks with your notes in one place. Add tags to bookmarks for easier search and revision. Create bookmarks and notes in your browser extension and manage your library on the web.

Want to share interesting sites you found? Just create a group, invite your friends or colleagues and have all shared bookmarks in one shared place. You can also share your thoughts of a bookmark through annotations. Get the brainstorming done in no time.

Insert annotations into web pages and into PDF documents with the same tool.

Annota comes with Firefox extension that lets you create and share bookmarks, insert tags, highlight text and comment it while you browse the web.

You can find more information in About section.

Easy guide how to use Annota:

brUMo

Try out also brUMo - Browser-based User Modelling and Personalisation Framework.

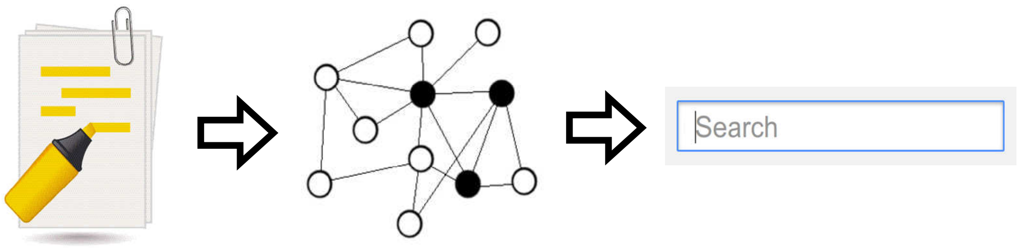
Figure 1 Annota, insert annotations into documents

II. Method

Query construction for related document search:

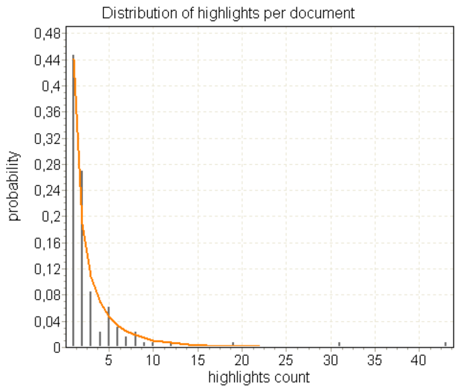
1. Text to graph transformation
2. Spreading activation

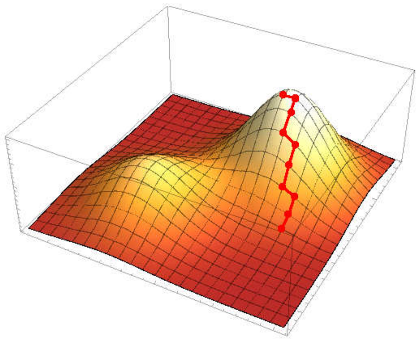
Text words transformed to graph nodes. Edges created using neighborhood of words. Initial activation inserted to nodes with attached annotations. Spreading activation in created graph. Nodes with concentrated activation used as query words. We are using different weights for various annotation types. Important is number of iteration.



III. Simulation

We have preformed a simulation to determine best weights for different types of annotations and to compare with commonly used TF-IDF based method. We analyzed properties of annotations from Annota.





We have generated annotations into documents from Wikipedia using parameters of annotations from Annota. Using proposed method, we created a query from document content and annotations. We have searched for similar documents. Relevant documents were those from the same category. We used hill climbing algorithm to optimize query construction parameters. Comparison of spreading activation based and TF-IDF based methods using simulation is in Table 1.

Table 1 Results of simulation with query construction using generated annotations

Method	Precision
TF-IDF, no annotations	21.32%
Proposed, no annotations	21.96%
TF-IDF, generated annotations	33.64%
Proposed, generated annotations	37.07%
TF-IDF, whole fragment annotated	43.20%
Proposed, whole fragment annotated	53.34%

Annotations are improving query construction process for different keyword extraction methods. Proposed method outperformed TF-IDF based method. Query construction process is independent from other documents but can be easily extended.

IV. References

[1] Faculty of Informatics and Information Technologies, Retrieved from http://www.fiit.stuba.sk/

[2] PeWe (Personalized Web) Group, Retrieved from http://pewe.fiit.stuba.sk/

[3] Annota, Retrieved from http://annota.fiit.stuba.sk/