



Insuma*Scout* Specification

Insuma GmbH

October 25, 2005

Insuma reserves the right to make changes in the specifications and other information contained in this publication without prior notice. In case of doubt, the reader should consult Insuma to determine whether any such changes have been made. The software described in this document is furnished under a license and may be used or copied in accordance with the terms of such license.

Trademarks

Insuma and its logo are registered trademarks of Insuma GmbH. All other product and company names are either trademarks or registered trademarks of their respective companies.

Insuma GmbH
Sand 13
72076 Tübingen
Germany
Phone: +49 (7071) 3 65 96 10
Fax: +49 (7071) 3 65 96 14
E-mail: info@insuma.de
<http://www.insuma.de>

Contents

1	Introduction	5
2	Features	6
2.1	Web-based collection creation	6
2.2	Delivery of new documents	6
3	Product specification InsumaScout	7
3.1	Web page archiving	7
3.2	Improved document similarity detection	7
4	Limitations and conclusion	8
5	Workflow	9
6	Lifecycle of a document	10
7	Lifecycle of the system	11
8	Working cycle of the system	12
9	Functionality of the Control Center	13
10	Typical usage scenario	14

Executive summary

This document describes the *InsumaScout* product, a part of *InsumaFocus* search engine for Web portals and intranets. The overall functionality is described as well as separate plugins and their function.

1 Introduction

The *InsumaScout* product is a set of tools for discovery, ranking, classification, and subsequent retrieval of Web documents relating to predefined topics, herein called “topic collections“ or just “collection“. A collection is defined by the collection administrator through core documents, i.e. a set of documents or its excerpts, plus start URLs, which are presumed to contain relevant data. The document ’s excerpts may be given by URLs pointing to corresponding publically available documents. Information is gathered automatically by Web harvesting robots (called Crawlers in the current document). The found documents are supposed to be useful for the collection’s topic defined by the user.

The client enters the following initial data for a topic:

- 10 to 20 paragraphs of highly relevant texts per topic and URLs which point to publically available documents or excerpts of documents which provide highly relevant text to the topic.
- Start URLs: 20 to 30 URLs pointing to the Web sites containing at least some relevant pages.

This information is used by the Web harvester (crawler). Please note: The more relevant information given, the better can the crawler filter be trained to fetch related topics from the Web. The crawler is configured to present a maximum of 100 documents per collection and week for the attention of the client ’s personnel or external experts. The newly fetched documents will appear every Wednesday during the *InsumaScout* licence period. A Web-based interface (also called “*InsumaScout* control center“) will allow the client ’s personnel to review the presented documents and rate their relevance using the scale shown in Table 1.

Table 1. Relevance levels

-2	Totally irrelevant
-1	Not relevant enough
0	No judgement
1	Relevant
2	Very relevant

The ranking information provided will be used for automatic improvement of the filters used by the Crawler. A process of continuous automatic learning will thus be sustained. The positively ranked documents will be added to the searchable document database suitable for subsequent retrieval.

2 Features

2.1 Web-based collection creation

All operations needed to set up a new topic, start and stop a crawler, change an existing topic, etc. will be available via an easy to use Web-based interface, called *InsumaScout* control center. This will minimise the system administration overhead.

2.2 Delivery of new documents

This feature will allow the user to target the Crawler for detection and delivery of only the new documents (documents recently posted to the web or documents recently modified). HTTP protocol can provide document modification dates. Still, in a large fractions of web servers this information is not provided. To identify the document as new, the Crawler will use a modification database and a tracking database.

3 Product specification InsumaScout

Modification database will keep the dates of review and characteristics of the document for the reviewed documents. Additional opportunity to define the modification or creation time can be provided by recognition of semi-structured data (e.g., the creation date appearing as plain text in the body of a web page).

3.1 Web page archiving

The production system will most probably fetch much more documents than it is feasible to review manually. Only a small portion of the documents that pass the filter will be presented for the expert review. The other will end up in a document database. The users will be able to retrieve them from the database using an advanced search interface supporting searches by keyword, category, modification date, and link structure. The software supports collaborative annotation and categorisation of documents. The ability to use a unique identifier for each version of each harvested Web page is valuable to writers of internal reports since they will be able to cite Web sources. By doing caching of documents in the database the available local copies of documents speed up browsing and save bandwidth.

3.2 Improved document similarity detection

Often several almost identical copies of a document co-exist located at different addresses, or a document does change only technically between two crawls. While the pilot system has placed such documents into the same cluster if they were fetched during the same weekly feed, this is not sufficient. Such documents must be treated as duplicates and must be presented only once. To achieve this, a lingua-statistical comparison of suspicious document pairs will be undertaken to estimate their similarity.

4 Limitations and conclusion

The customised product proves the applicability of the *InsumaScout* technology to the client's needs for Web information discovery and classification. The client's personal or external experts can provide expert knowledge in forms of document rates. Those re-adjustments will gradually improve and train the overall system with respect to the information fetched from the Web.

5 Workflow

InsumaScout allows for automatic discovery of relevant documents from the Internet, their management, classification, and re-use. The major scenarios supported:

- The user setups starting URLs and gives an example of a relevant text. The system collects relevant documents.
- The user rates the documents in the control centre, in gradation of relevance. The crawler uses this feedback to improve the relevance of newly fetched documents.
- The user assigns documents to a (hierarchical) tree, the system automatically assigns all new documents to these categories (routing).
- The user searches among the documents.
- The user integrates the search into own portals using XML interface.

6 Lifecycle of a document

The lifecycle of every document in the system looks like follows:

- Relevant document in Internet is discovered by thematic crawler (see Core text and Start URLs in the Crawler settings)
- Discovered documents are grouped by fetch rounds (browse by Fetch rounds)
- Documents can be reviewed and rated
- Documents in the collection can be searched by keywords (Search)
- Rating influences the crawler filter and improves relevance of documents delivered in the next fetch rounds

7 Lifecycle of the system

There are 2 major periods of the system usage:

- Training period
- Routine period

In the training period (5-10 fetches, or a 1-2 months) the user should invest some effort into ranking of delivered documents, so the system can learn the relevance requirement. As soon as the system starts delivering satisfactory relevant documents, no further ranking necessary. The documents continue to be added, the knowledge base grows on its own.

8 Working cycle of the system

The system runs in so called **fetch rounds**. In the first fetch round the user provides starting URLs, samples of the relevant text and starts the crawler by generating an initial filter. The crawler estimates relevance of a particular page against this filter. It can be manually adjusted by deleting not relevant words. The user can "kill" terms that seem to be not appropriate for the topic (too generic, wrong language, etc.) In the second and further fetch rounds the user rates delivered documents in order to adjust the crawler filter more precisely to the topic. Fetch round takes usually 1 week, with new documents appearing on the system on Wednesday.

A fetch round includes following steps:

- Crawler is running and collects the documents
- The documents appear in the the control center
- Admin can browse through the documents and rate them
- The collection can be browsed and docs can be re-rated
- The crawler filter is beeing updated based on user rating

9 Functionality of the Control Center

- See the documents discovered by crawler grouped by fetch rounds
- Browse the documents fetched in the last fetch round and rate them
- Search the entire collection
- Use tools for XML integration in your portal

10 Typical usage scenario

A typical usage scenario would be:

- The client company activates *InsumaScout* in order to provide thematic portal to it's users
- Technical staff at client company enters sample relevant documents under Crawler options > Relevant text samples
- Starting URLs as well as allowed URLs are entered under Crawler settings
- An initial filter based on relevant samples is generated Crawler options > Train filter
- The generated filter can be adjusted manually in order to improve its discriminating ability
- Crawler brings new documents, a new fetch appears under Crawler > Fetches as scheduled (normally every Wednesday)
- Knowledge experts login and rate the delivered documents, from -2 (irrelevant) to +2 (very relevant).
- The crawler corrects the filter and performs the next fetch
- When the newly fetched documents become more or less relevant, the exploitation or (routine) phase may begin.
- The users at the company can search the collected knowledge base under Indexer
- The technical staff of the client company integrates the search into their portal, using XML interface and examples, provided in the Technical Library.