User-Centric Search over Multimodal and Multimedia Content

Gregory Grefenstette[1], Petros Daras[2], Efstratios Tzoannos[3], Vincenzo Croce[4], Jonas Etzold[5], Vasilis Tountopoulos[6], Alberto Massari[7], Sabine Spiller[8], Lorenzo Franco Sutton[9]

[1]Exalead, Paris, France; [2]CERTH/ITI, Thessaloniki, Greece; [3,6]ATC S.A, Athens, Greece; [4]Engineering Ingegneria Informatica S.p.a, Italy; [5]University of Applied Sciences Erfurt, Germany ; [7]University of Genoa, Italy; [8]EasternGraphics, Germany; [9]ANSC, Rome, Italy

E-mail: [1]ggrefens@exalead.com, [2]daras@iti.gr, [3]e.tzoannos@atc.gr, [4]vincenzo.croce@eng.it, [5]jonas.etzol@fh-erfurt.de, [6]v.tountopoulos@atc.gr, [7]i.search@infomus.org, [8]Sabine.Spiller@EasternGraphics.com, [9]l.sutton@santacecilia.it

Abstract: **Beyond current textual search, there remains a need for greater variety in query modality and in media input for querying. The I-SEARCH project aims at developing an experimental platform for new types of querying of multimedia document sources. This article presents an overview of the I-SEARCH project, shows some typical use case scenarios for the future I-SEARCH platform, and describes the architecture that we have designed for implementing these platforms.**

Keywords: search, multimodality, multimedia, content-based indexing, user-centric

1 INTRODUCTION

As the digital world moves from textual content (with its nonetheless rich variety of formats) into a world of greater availability and greater demand for multimedia content (2D images, drawings, video, 3D objects and audio), we are faced with the problem of how to represent these disparate media in a common way so that information can searched no matter what media carries it. And, as users become more used to dealing with various media as output, there will be a growing demand to expand search queries from simple typing of text into an multimodal expression that the user feels would describe their information need, i.e., showing an image, speaking, drawing a shape in the air, capturing a video. We have begun a multiyear project, called I-SEARCH, studying next generation search. This search will be highly user-centric delivering the relevant (though heterogeneous) content, introducing novel visualisation schemes to optimally present results to users, results which are dynamically adapted to the end-user's device, be it a simple mobile phone, a computer pad, a laptop, or a high-performance PC.

2 I-SEARCH FUNCTIONALITY

I-SEARCH is principally concerned with expanding the breadth of what is typically indexed in current search engines. Figure 1 shows the typical phases of a search application, with indexing on the right and retrieval of results on the left. Analyzing input content, no matter what its input media or real world context, is handled in the Content Analytics step which converts raw signal into structured content that can subsequently be indexed. In its approach to Content Analytics, I-SEARCH addresses the challenge of dealing with the different modalities of the multimedia items. In this phase, media and modal-specific analysis will be performed to extract structured multi-level descriptors. This structure includes: i) content specific low-level descriptors, which can characterise the type of content, ii) real-world descriptors, which associate the content with information extracted from sensors (i.e. GPS, temperature, time, weather, RFID information, etc.), and iii) user-related descriptors, which encapsulate expressive, social and emotional characteristics to the semantics of these items.

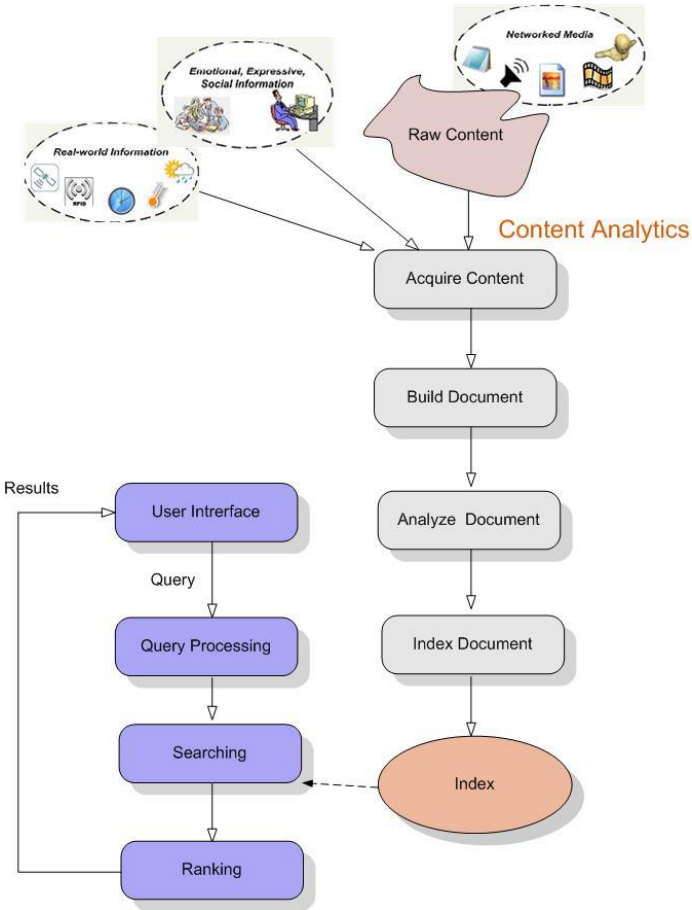

Figure 1: Typical Phases of a Search Application. I-SEARCH plans to capture social and real world information at the same time as the raw text and multimedia content

Corresponding author: Gregory Grefenstette, Exalead, 10 place de la Madeleine, 75008 Paris, France, ggrefens@exalead.com

In order to provide this structural description of the metadata for the multimodal and multimedia content, I-SEARCH defines the Rich Unified Content Description (RUCoD) formulation. RUCoD will integrate features coming from the above-mentioned descriptors, providing a standardised way to annotate multimedia content. The definition of RUCoD comprises the core idea of the I-SEARCH project and will drive the specifications of all the components involved in the Content analytics phase of the I-SEARCH search platform. The development of the RUCoD format is still undergoing refinements but Figure 2 gives the general structure.

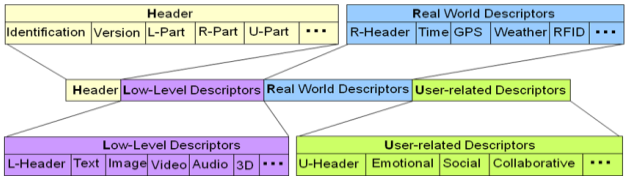

Figure 2: The RUCoD General format

RUCoD records are composed of a header, containing general information about the content object (type, name, ID, creation information, and general information about the different media: 3D, images, sounds, videos, text). This header is followed by the detailed descriptors themselves: low-level descriptors, compatible with MPEG standards [1], extracted from each separate media (3D, images, sounds, videos, text); real-world descriptors [2][3] representing time, weather, location, etc.; and descriptors related to the user behaviour (emotions, expressions) [4][5] are stored.

I-SEARCH will also provide novel techniques for multimodal annotation propagation [6], which will be used to provide a complete description of the semantics of an item in the RUCoD representation even though this information may not have been available in the original content. Missing data will be inferred via exploitation of similarity functions for known multimodal and multimedia objects [7].

3 TYPICAL USE SCENARIOS

I-SEARCH seeks to provide new multimedia and new multimodal descriptors that can be search without regard to mode or media. This novelty can be best seen in some of the I-SEARCH use cases.

3.1 Music retrieval through expressive embodied queries

Modern search engines include non textual querying modalities such as audio (Shazam [8], Google China Music) or image (Google Goggles, Google similar images). These query-by-content or query-by-example engines are useful when the research objective is clearly defined or where specific information is targeted. Within the I-SEARCH framework, alternative yet complementary querying modalities are added (e.g. expressive gesture, affective and emotional cues) that facilitate more explorative and creative search.

Suppose a user is in an environment equipped with devices [8][10][10] which support multimodal search and retrieval through the I-SEARCH platform, allowing the user to express

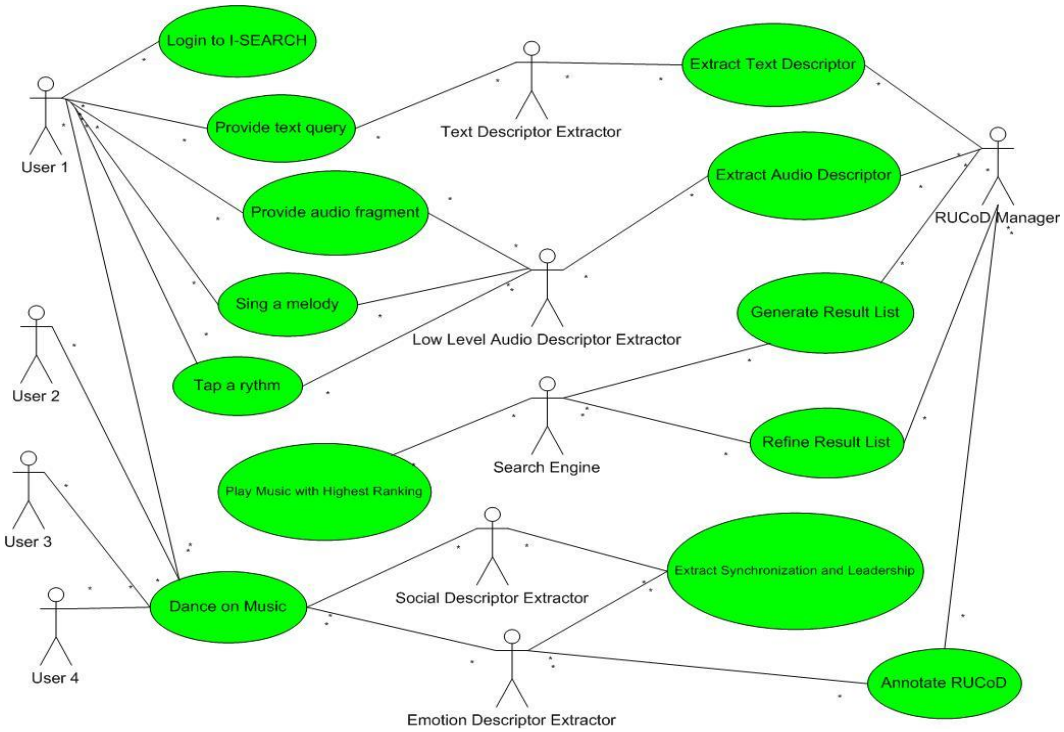

Figure 3 Music retrieval scenarios using multimodal search and filtering via Rich Unified Content Descriptors.

themselves through voice or hands/body movements. Pre-recorded multimedia content can be uploaded from external device (such as mobile phone) through wire or wireless connections (e.g., bluetooth) [11][12]. This interactive environment is connected to the I-SEARCH platform, which interfaces with a server located at a digital library. The contents of the digital library collection have been pre-processed using I-SEARCH platform components to extract RUCoD descriptors related to low-level features, real-world data and expressive/ emotional/ social cues. Now consider the following scenario:

1. Chiara wants to explore music artworks that share affective features with the Ravel's Bolero, using the e I-SEARCH framework to retrieve audio information that share similarities with this audio pattern. After connecting to the I-SEARCH homepage on her computer, she chooses the audio icon and selects beat capture to enable the audio recognition functionalities.
2. Chiara uses the recording she previously made with her cell phone to submit an audio query to the I-SEARCH engine.
3. The I-SEARCH framework extracts low-level descriptors from Chiara's supplied audio content and creates a new, internal query based on the RUCoD format. Similar audio results related to Bolero are retrieved. Related video files, images and music scores are also retrieved through the multimodal annotation propagation, or automatic detection of common properties like real-world location (e.g. photographs with GPS geo-tagging information that place the subject in the same location where a specific Bolero recording was captured). Results are displayed via visual analytics techniques on Chiara's computer screen, using clusters annotated with information like modality type, population size, etc.
4. Chiara picks one of the results returned by the query, listens to it but decides that what she needs is something more energetic
5. She closes her fists and starts making sharp, sudden vertical movements on the same Bolero rhythm. Through a video camera or embedded accelerometers the environment captures the emotional features of the gesture and changes the displayed results to match it (either by removing the items that don't convey that emotion, or by moving the suitable items closer to build a different cluster configuration)
6. One of the results captures Chiara's attention: a drum recording from Italian ethnomusicological repertoire, the 'Ritmo di tamburo' from Raccolta 24 (eg. Piece 119 at around 00:30, piece 124 at around 00:08) where various drum rhythms are played, sharing indeed the same triplet of Bolero. A little further away she also finds a voice recording.

A diagram that covers this scenario is shown in 3.

3.2 Search for a Piece of Furniture

Currently, in the area of furniture data modelling and retrieval, explicit Product Catalogues are used, which, apart from commercial and textual information, also contain 3D and 2D (ground-plan perspective) data models and (static) images (high-quality renderings and photos). Now suppose an end user aims to search for a particular article of the office furniture industry, such as a chair, a desk or some container. The input for the search can be an image (e. g. photo, sketch, rendering of a 3D model), some text (e.g. words, phrases, letters) or a 3D object (e. g. CAD design of a piece of furniture). Consider this scenario:

1. A user enters the I-SEARCH interface to perform a product catalogs search
2. The end user provides an image, text or 3D input to the RUCoD framework (e. g. photo, sketch, rendering of a 3D, text, 3D object)
3. The I-SEARCH platform extracts low-level descriptors from the input, and creates a query based on the RUCoD format.
4. The user can also include social and real-world

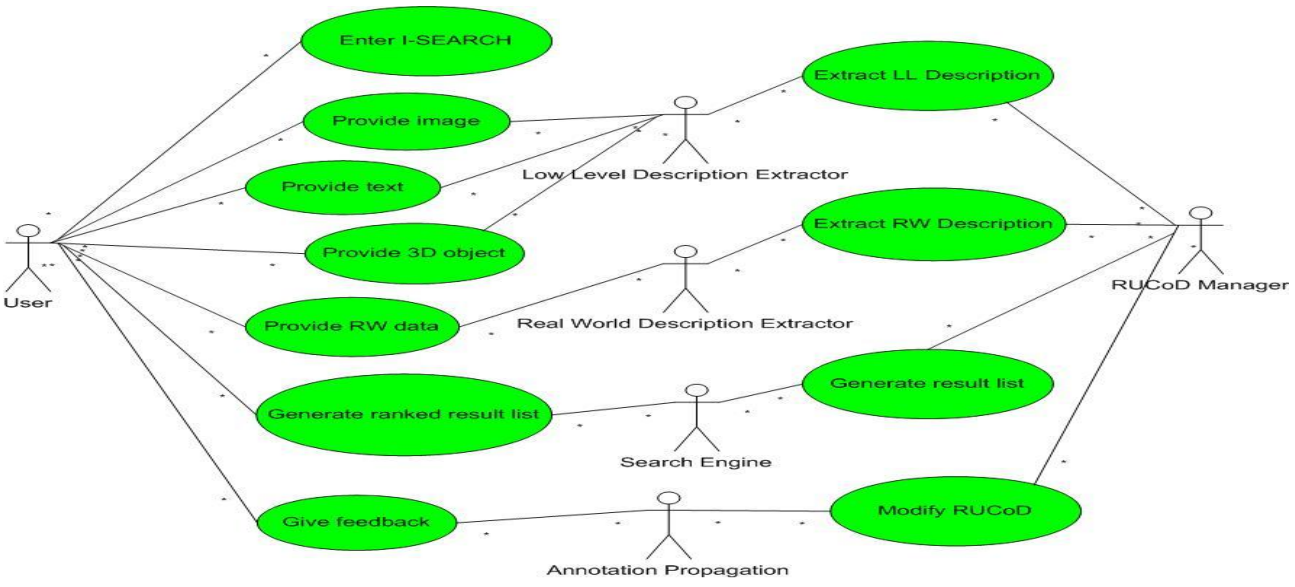

Figure 4 Search and retrieval use cases for 3D furniture models

descriptors to the search procedure. For example the end user could look for an article, which is best evaluated by former customers (social descriptor), or/and, where a dealer is available within a radius of 50 km (GPS, real-world data), or/and, which is not older than 10 years (temporal information, real-world data), or/and, which is described by intrinsic product properties

5. The retrieved results are presented according to the chosen preferences of the end user as well as to their similarity/matching with the query input. The I-SEARCH platform now displays the results as groups (e.g. created by the modality of the search results), where the most characteristic object based on the search query is selected as the representative.
6. If the end user is not satisfied with the retrieval result, relevance feedback or interaction with the content could be applied to get a more refined output.

A diagram that covers this scenario is shown in Figure 4 above.

4 ARCHITECTURE FOR I-SEARCH

As it can be seen from the scenario figures, the I-SEARCH platform is designed to allow end users to submit queries and get relevant results based on the content analysis producing multimedia tags as they are defined in the RUCoD specification. From a systems point of view, the use cases, along with other use cases not presented here, can be abstracted to a general use case dictating the general functionalities that the I-SEARCH platform should perform. The relevant Use Case UML (Unified Modeling Language) diagram [13] of the system is depicted in Figure 5. This diagram represents the use cases for each user group and the relationship between the actions and the users. Following the search application phases shown in Figure 1, Figure 5 details the main actions that will be performed by I-SEARCH platform users:

- On the one hand, administrative users will be involved in the analysis phase to reformulate the target multimodal content, so that the content can be made accessible to the end users performing their search and retrieval tasks. As mentioned before, RUCoD specification to be defined and refined in the project will be used to describe the content in the appropriate format.
- On the other hand, the end users are involved in the core search and retrieval phase of the I-SEARCH platform and invoke the relevant system functionalities, in order to formulate their queries, based on the desired search modality, and provide their input on the retrieved results list in order to refine their search.

4.1 Architecture Design Methodology

In order to derive the architectural design of the I-SEARCH project in an R&D project that aims to deliver a platform prototype implementation to be tested and evaluated over at least three target domains, a rapid prototyping methodology [14] has been selected and adopted. This approach is adapted to the project parallel process between requirements gathering from the end users and developing the I-SEARCH platform, allowing early adopting business partners to actively participate in the specification of use cases and the evaluation

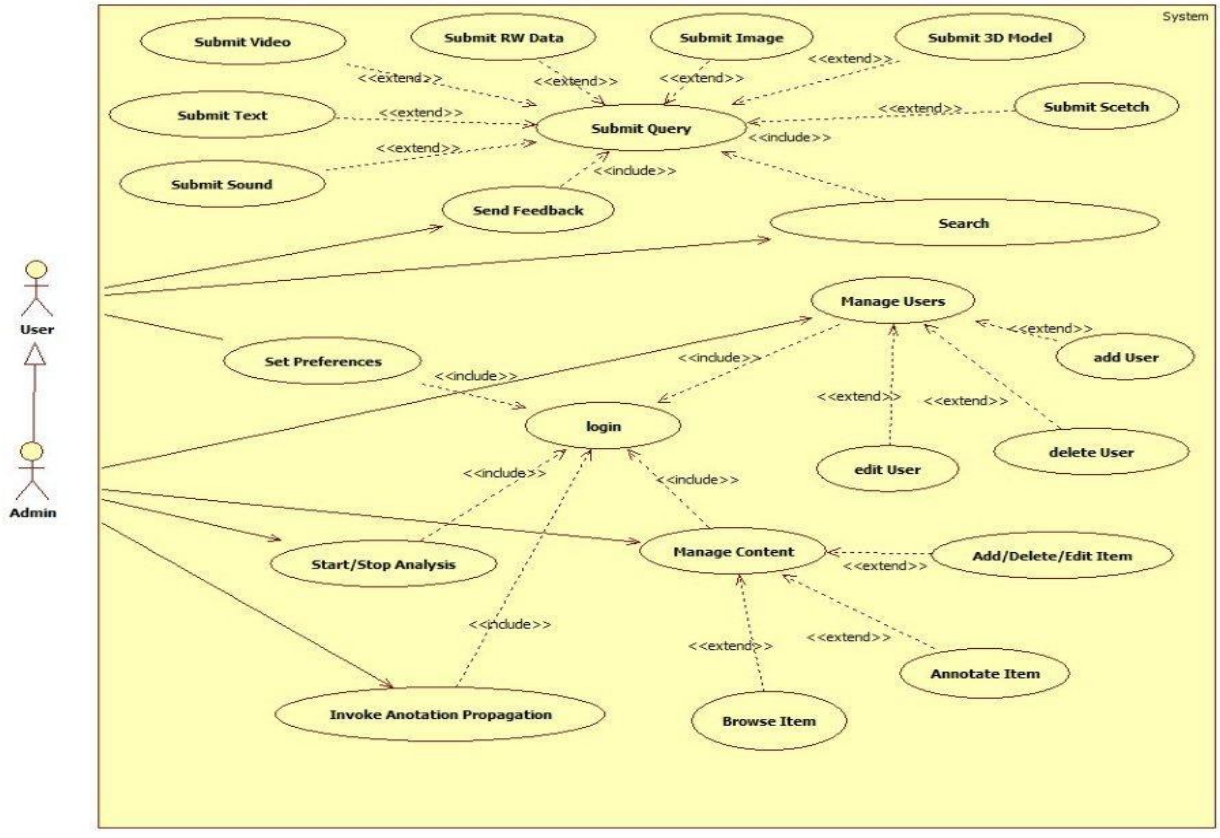

Figure 5 UML Diagram of I-SEARCH system

of the system developments and provide valuable feedback in an iterative way, as well as enabling us to release platform prototypes to the market in a punctual manner, in a time window that enables for attracting target users, before the technologies conveyed become obsolete and have no acceptance to the market.

4.2 High Level Overview of the Architecture

A conceptual diagram of the system outlines clearly the module interactions along with the user activities. In the Figure 6, the rectangles depict functional modules of the system and the round-edge boxes describe the actions taken by the user or the system. After the user logs in or is identified by a stored certificate or other authentication method the user's personalised information is loaded and then he or she is able to select the action to perform.

Provisioning, platform administration, and content feeding is accessible only to privileged users. The basic user will use the search and relevance feedback features of the system. However, unauthorised users can perform simple search requests to find the desired multimedia. It should be clarified that, in Figure 6, the dotted lines correspond to user actions, while the solid ones correspond to system flows.

5 CONCLUSION

I-SEARCH aims to provide a novel unified framework for multimodal content indexing, search and retrieval. This framework will be able to handle multimedia and content along with real world information and user related information, which can be used as queries and retrieve any available relevant content of any of the aforementioned types. The I-SEARCH search engine will be highly user-centric in two senses: (i) only the content of interest will be delivered to the end-users, satisfying their information needs and preferences, and (ii) the user can engage in physical, multimodal interaction with the system to refine search results. Furthermore, by introducing novel visualisation schemes, the retrieved results will be optimally presented to the user, which is expected to dramatically improve end-user experience. Finally, the search engine results will be dynamically adapted to end-user's interaction devices, which will vary from a simple mobile phone to a high-performance PC.

This article provides the early-stages outline of how the I-SEARCH platform architecture has been designed and will be implemented and developed over the coming years. The architecture design described here will be updated during the project, and we welcome any comments by interested parties.

Acknowledgments

The I-SEARCH project is funded by the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 248296.

References

[1] Martinez JM., Koenen R., Pereira F. "MPEG-7: The generic multimedia

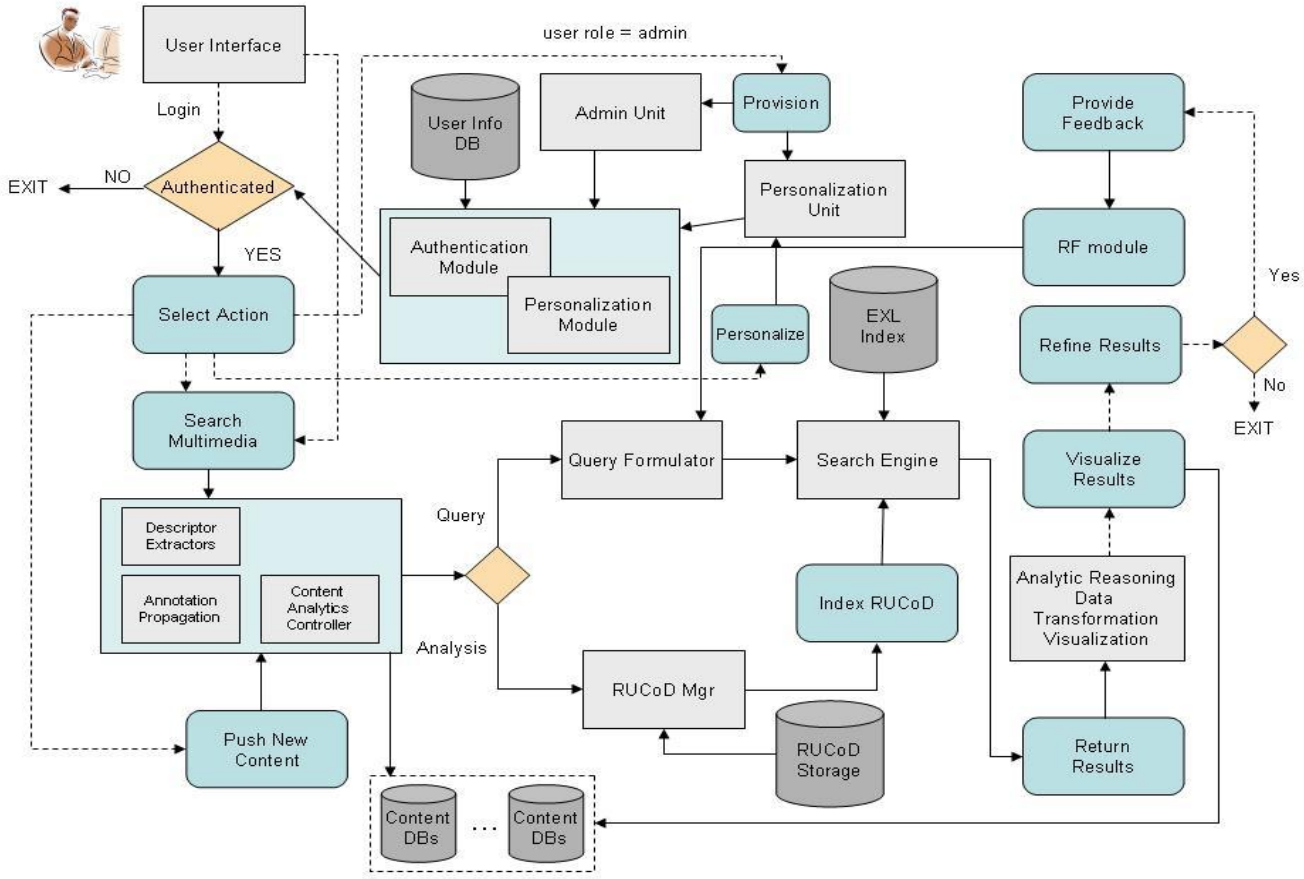

Figure 6 The I-SEARCH Flow Chart

content description standard, part 1". IEEE Multimedia 9(2):78–87, 2002.
[2] Camurri A., Coletta P., Demurtas M., Peri M., Ricci A., Sagoleo R., Simonetti M., Varni G., and Volpe G. "A Platform for Real-Time Multimodal Processing", in Proceedings International Conference Sound and Music Computing 2007 (SMC2007) Lefkada, Greece, July 2007.
[3] Yick, J., Mukherjee B., Ghosal D. Wireless sensor network survey. Computer Networks, 52(12):2292–2330, 2008.
[4] Frisson C. et al. "Bodily Benchmark: Gestural/Physiological Analysis by Remote/Wearable Sensing". In: QPSR of the numediart research program. Ed. By Thierry Dutoit and Benoît Macq. Vol. 2. 2. Numediart Research Program on Digital Art Technologies. pp 41-57, 2009.
[5] Wagner J, André E, Jung F. "Smart sensor integration: A framework for multimodal emotion recognition in real-time". In: Affective Computing and Intelligent Interaction (ACII 2009). 2009.
[6] Agosti, M. and Ferro, N. "A Formal Model of Annotations of Digital Content", ACM Transactions on Information Systems, 26(1) Nov 2007.
[7] Town C. "Ontological inference for image and video analysis," Mach Vision Appl, vol. 17, no. 2, pp. 94–115, 2006
[8] Wang A. The Shazam music recognition service. Communications of the ACM, 49(8):48, 2006.
[9] Poppe R. "Vision-based human motion analysis: an overview." Computer Vision and Image Understanding, 108:4–18, 2007
[10] Aylward R, Paradiso JA. "Sensemble: a wireless, compact, multi-user sensor system for interactive dance." In Proceedings of the 6th Conference on New interfaces for Musical Expression (NIME '06), pages 134–139, 2006.
[11] Camurri A, Bevilacqua F, Bresin R, Maestre E, Penttinen H, Seppanen J, Valimaki V, Volpe G, Warusfel 0. "Embodied music listening and making in context-aware mobile applications: the EU-ICT SAME project". Proceedings of ICMPC 10, 67:3969–3972, 2008.
[12] Casey MA, Veltkamp R, Goto M, Leman M, Rhodes C, Slaney M. Content-based music information retrieval: current directions and future challenges. PROCEEDINGS-IEEE, 96(4):668, 2008
[13] Fowler M. UML Distilled: A Brief Guide to the Standard Object Modeling Language. Second Edition, Adisson-Wesley, 1999
[14] Jones TS, Richey RC. "Rapid prototyping methodology in action: A developmental study." Educational Technology Research and Development, *48*(2), 63–80, 2000.