

Cluuz Semantic Platform

For Better Information Understanding



“Yahoo’s search index is more useful to me with the Cluuz.com functions than those created by Yahoo’s wizards”

**– Stephen Arnold
Beyond Search**

Cluuz platform enables unstructured information processing through our patent-pending process of entity extraction, disambiguation, matching, clustering, and generation of semantic answers.

When we refer to unstructured information we mean information that is found in documents, web pages (html) and other sources of information where there has not been a specific pre-determined structure in terms of meaning of the different pieces of information within the document. To be more precise, we are talking about meaning that is machine understandable – while a document might have a heading that is understood by humans, the heading is not interpreted by normal software processing activities (such as displaying a web page) any different as far as the semantics of the heading is concerned. The markup in web pages is used for display purposes. In recent years there has been an advent of the semantic web, to the point of semantic becoming part of the HTML5 specification, through the inclusion of microdata, which should enable sites to publish semantically structured information within HTML documents. However, the effort required to markup data semantically is still quite substantive. Almost all of the existing information on the web is unstructured, and most of the future information will be unstructured for a foreseeable future.

Cluuz processes unstructured information through natural language processing techniques to bring new level of understanding of information based on semantic concepts. The processing is done through a set of steps from information element identification, normalization, entity extractions, disambiguation, clustering and generation of semantic answers.

Information Identification

Since most of the data that Cluuz deals with is web pages, we will concentrate on the processing of unstructured information found in web pages, however, it should be noted that Cluuz deals with Office documents, pdfs, regular text and other document formats, as long as they accessible either on the web or locally. Once Cluuz encounters a web page first step is to isolate parts of the web page that are information rich. For example, a typical web page has navigational and advertising elements that are not information rich – and main content of the

Cluuz Semantic Platform

For Better Information Understanding

page that does contain information of interest. Cluuz evaluates different page elements through a set of algorithms that look at page structure, markup-to-text ratio and other metrics to find information rich sections. At this stage, Cluuz also identifies image elements in the page, and identifies images that are considered to be relevant to information. Navigational and advertising images are discarded.

Normalization

In this step Cluuz takes the information rich HTML content and converts it into semi-structured information. The HTML tags such as headers and paragraphs are used as means of creating an overall document structure, similar to table of contents, where the large portions of text are associated with an entry within the structure. Note that this process is also done for non-html document sources. For HTML documents, the non-semantic markup (standard html tags) are removed and the text is used as elements of the document structure. This process therefore results in a normalized representation of the document with headers/chapters and paragraphs within chapters. Further to this step, each of the paragraphs is broken down into sentences, and finally tokens.

Entity Extraction

Given the normalized document, Cluuz is able to perform entity extraction. The types of entities that are extracted currently are: People, Companies, Organizations, Phone Numbers, Email Addresses, Street Addresses, Domains, Dates, Geography. The catch-all for the entities is the Other category. The entities are extracted out of the textual structures based on complex set of rules that take into account position of token in the sentence, part-of-speech tagging, regular expressions, dictionary lookups, predicates, prefixes, suffixes and other techniques. Each entity is associated with a category and extended entity information. The entities are stored within the chapter/paragraph/sentence structure and where possible entity relationships are identified. Entities are forward propagated within paragraphs, which deals with situations such as defining the entity at the top of the document and then referring to the entity further down in the document (e.g. defendant/claimant ..). Entities are therefore associated with other entities through either direct relationships, sentence association, paragraph association, chapter association or document association. This graph is what we call documents semantic graph.

Cluuz Semantic Platform

For Better Information Understanding

Entity Disambiguation

“A search for “*Kate Greene*” immediately pulls up my e-mail address at Technology Review, the university I attended, and a number of the people I’ve interviewed for past stories. Additionally, Cluuz provides other tools that allow the links and relationships between different semantic concepts to be visualized easily.”

**– Kate Green
MIT Technology
Review**

Given a set of documents, and their associated semantic graphs, we can now look at different entities extracted from different documents and see if they are the same entities and consequently know if the documents are talking about the same entities/topics. The process consists of comparing the semantic graphs and looking for significant overlaps. This process is akin to the mental process that we, humans, go through when being asked if we know someone – if we were asked about someone with ambiguous name (for example, lets assume that there are two John Smith's that we know) we would ask for further details, such as place of work, address, companies that they are involved in and using this additional criteria we would know which particular John Smith we are talking about. Cluuz algorithms perform similar process in order to disambiguate entities. Once an entity is found that is the same within multiple semantic graphs, one of the entities is removed and graphs are joined through the common entity. Hence, the entity disambiguation results in a minimal set of isolated semantic graphs which are clusters containing references to one or more documents.

Clustering

Given a set of clusters, generated as a result of entity disambiguation, we can now analyze the clusters to understand their nature. For example, Cluuz currently looks at each of the semantic clusters to see what are the entities that are central to the cluster. The entity with the highest degree centrality is considered top entity for the cluster. In the cases where the graphs are generated as result of a search, the entities within the query are not considered candidates for the top linked entities because it is assumed that all the document have relationship to these entities since the search results should be relevant to the query and instead other most central entities are used as top linked entities. The clusters could be analyzed for other features such as most common terms, which could be used as another indicator of the nature of the cluster.

Cluuz Semantic Platform

For Better Information Understanding

Semantic Answers

Given a set of documents that are retrieved as a result of some query, Cluuz is able to find what it deems the most appropriate answer by evaluating each one of the sentences in the plurality of documents based on Cluuz semantic relevance metric. The relevance metric is calculated by evaluating entities and words within each sentence and determining if they are contributing to the possible answer to the given question. The question is analyzed to determine the intent, and given an intent, different entity types are given different weight as to how they contribute to the score. For example, if a query is interpreted as asking for a person, sentences with person names are given higher rank than sentences without person names. In addition, the keywords within the query are identified and synonyms and other related terms based on language ontology are identified which also contribute to the score.

Availability

Cluuz algorithms are accessible through online web service calls as well as an embeddable API which can be used within applications. To see Cluuz in action please go to Cluuz.com and perform some searches – note that none of the extraction is done ahead of time, meaning that all Cluuz algorithms are running on information obtained in real-time. Most of the response time is spent in awaiting data from origin sites – Cluuz processing time represents in most cases but a few percent of the response time.



Sprylogics Inc. 55 Administration Rd Unit 12, Concord, Ontario, L4K 4G9

© Copyright 2011, Sprylogics Inc.

Cluuz Semantic Platform

For Better Information Understanding