PUC

ISSN 0103-9741

Monografias em Ciência da Computação
n° 01/14

Handling Google Snippets with SWI-Prolog

Edirlei S. de Lima
Antonio L. Furtado

Departamento de Informática

PONTIFÍCIA UNIVERSIDADE CATÓLICA DO RIO DE JANEIRO
RUA MARQUÊS DE SÃO VICENTE, 225 - CEP 22451-900
RIO DE JANEIRO - BRASIL

Monografias em Ciência da Computação, No. 01/14 ISSN 0103-9741
Editor: Prof. Carlos José Pereira de Lucena April, 2014

Handling Google Snippets with SWI-Prolog

Edirlei S. de Lima
Antonio L. Furtado

{elima, furtado}@inf.puc-rio.br

Abstract: We have designed – and implemented in a preliminary version – a tool, named **LOG-SNIP**, for capturing snippets while performing Google searches for Web resources pertaining to a domain of interest, based on keywords adequate to delimit the domain. The snippets are decomposed into separate fields: name, date, url, info. A kws field is added by extracting resource-specific keywords from the name and info fields. Under the form of a five-field frame structure, the chosen snippets can then be recorded as Prolog clauses, to be subsequently used for all sorts of research purposes. Of particular value is the ability to employ the sets of resource-specific keywords to perform comparisons among the located domain resources. To present one possible application, we implemented a module that translates the stored clauses into the clauses required to run our previously created **KW-GPS** tool.

Keywords: Google Snippets, Web Resources, Keyword Search, Logic Programming.

Resumo: Projetamos – e implementamos em versão preliminar – uma ferramenta, chamada **LOG-SNIP**, para capturar "snippets" no decorrer de buscas via Google por recursos na Web pertencentes a um domínio de interesse, baseadas em palavras-chave adequadas para delimitar o domínio. Os "snippets" são decompostos em campos separados: name, date, url, info. Um campo adicional kws é produzido pela extração de palavras-chave específicas do recurso a partir dos campos name e info. Sob a forma de uma estrutura de "frame" com cinco campos, os "snippets" escolhidos podem então ser registrados como cláusulas Prolog, a serem depois utilizadas para qualquer finalidade de pesquisa. Particularmente valiosa é a capacidade de empregar os conjuntos de palavras-chave específicas para realizar comparações entre os recursos do domínio localizados pela busca. Para apresentar uma possível aplicação, implementamos um módulo que traduz as cláusulas armazenadas nas cláusulas que são requeridas para rodar nossa ferramenta **KW-GPS**, criada anteriormente.

Palavras-chave: Google Snippets, Recursos na Web, Busca por Palavras-chave, Programação em Lógica.

In charge of publications

Rosane Teles Lins Castilho
Assessoria de Biblioteca, Documentação e Informação
PUC-Rio Departamento de Informática
Rua Marquês de São Vicente, 225 - Gávea
22451-900 Rio de Janeiro RJ Brasil
Tel. +55 21 3527-1516 Fax: +55 21 3527-1530
E-mail: bib-di@inf.puc-rio.br

1. Introduction

We start from the assumption that the snippets exhibited as the result of a Google search, based on keywords chosen so as to define the domain of current interest, are sufficiently informative in a fair number of cases. At the very least, what they convey is often informative enough at a first stage, obviating the need to immediately engage in time-consuming access to each of the resources located. This is especially helpful if the resources are formatted according to the Google recommendations towards rich snippets and structured data.[1] One may also expect that enhancements in future versions of Google [Ribeiro-Neto] will result in snippets even more precisely tuned to the intended domain.

With this assumption in mind, we designed and implemented in a first prototype version the **LOG-SNIP** tool, for capturing the snippets while performing Google searches based on keywords adequate to delimit the domain at hand.

The snippets are decomposed into fields: name, date, url, info. A kws field is added by extracting resource-specific keywords from the name and info fields. Under the form of a five-field frame structure, the chosen snippets can then be recorded as Prolog clauses, to be subsequently used, in autonomous systems equally programmed in Prolog, for all sorts of research purposes. The choice of Prolog was motivated by the recognized suitability (cf. [Bratko], for example) of the logic programming paradigm for Artificial Language applications.

Of particular value is the ability to employ the sets of resource-specific keywords, after they are divided into classes previously chosen by the user, to perform aspect-oriented comparisons among the resources from which they were extracted, as does our previously created **KW-GPS** tool [Lima]. For this objective, we implemented a module that translates the snippet-generated clauses of **LOG-SNIP** into the format required to serve as input to **KW-GPS**.

The text is organized as follows. Section 2 describes the **LOG-SNIP** tool, as well the main features of the current prototype implementation. Section 3 covers the exploitation of the generated snippet files, both directly and after their translation into the clause format needed to run the **KW-GPS** tool. In both sections, the domain of detective stories [Christie, Hessick, Todorov] serves as example. Section 4 contains concluding remarks.

2. The LOG-SNIP tool

2.1. Functionality

The tool was designed to run in a Prolog environment. To send queries to Google and access the source html pages of the response, two predicates are provided:

search(L) - The input parameter L is a list of keywords and search directives recognized by Google. As soon as the command is entered, the system asks whether or not to extract keywords from the snippets. Then, as a result, each snippet found by Google is displayed in a four-field (name, date, url, info) or five-field (name, date, url, info, kws) format. As an option, the url can be activated to open and thus visualize the resource.

[1] https://support.google.com/webmasters/answer/99170?hl=en

store_sn(D, L) - The input parameter D serves to denote the domain of the search, which is used to compose the name of the Prolog file that will store the results, whereas the L parameter is again a list of keywords and directives. Each snippet, always in five-field format, is displayed and – if the user so indicates – stored in the Prolog file.

For both predicates, whenever an option is offered to the user, typing y is taken as a positive reply; a negative reply can be expressed by n but also by simply hitting the enter key. Both predicates allow the interruption of the process by typing end.

As said above, the list L supplied as input parameter contains keywords and directives to drive the Google search. Our intention was to mimic to some extent the Advanced Search Google interface, still in a reasonably user-friendly notation. An important difference is our way to delimit the time-interval, which is supported by the Google machinery[2] but not explicitly advertised in at least part of their documentation.[3] Our notation and its rendering into Google terminology is illustrated in the examples below. Note that a keyword can be either a single word or a keyword-phrase (with intervening blanks). An auxiliary predicate is called from inside the two predicates to operate the translation:

prep_google(L,U) - where L is a list of keywords and directives and U the generated url to guide the Google search.

In our first example below, the and term means that the located resources must contain all the words summary, victim, crime, investigation. The exact phrase Poirot stories must also be present, whereas the word blog is excluded by the minus sign. If several terms were to be excluded, the notation 'not(k1,k2,...kn)' could have been used. The last component is a directive, limiting the search to the English language. The starting date was left unspecified, so a default was applied (the date when we did the search minus 10 years). The ending date remained unspecified, being understood as "today" (the day of the search) by the Google server.

```
:- prep_google(['and(summary,victim,crime,investigation)', 'Poirot stories',
   '-blog', 'in:english'], U).

U = https://www.google.com/search?as_q=summary+victim+crime+investigation&
    as_epq=Poirot+stories&as_eq=blog&lr=lang_en&tbs=cdr:1,cd_min:3/19/2004,
    cd_max: .

:- search(['and(summary,victim,crime,investigation)','Poirot stories',
           '-blog','in:english']).
want to see keywords?[y/n] y

name   : A Keyword-based Guide to Poirot Stories - PUC-Rio
date   : Jul 29, 2013
url    : ftp://ftp.inf.puc-rio.br/pub/docs/techreports/13_10_lima.pdf
info   : associated with the story, which include but are not limited to plot-
         summaries, narrative texts and .... kws(1, investigation, [clue:
         'character of the victim', snag: 'time of death', ... With the choices 7,
         12 for victim, 3, -11, 14 for crime, and 1, 2, 8 for.
keywords: [narrative texts, Poirot Stories, Keyword-based Guide,victim, kws, snag,
           PUC-Rio, clue, choices, story,plot-summaries, investigation, character,
           time, death, crime]
want to see?[y/n/end]:
```

[2] http://stenevang.wordpress.com/2013/02/22/google-search-url-request-parameters/

[3] https://developers.google.com/search-appliance/documentation/50/xml_reference
https://developers.google.com/custom-search/docs/xml_results?hl=en&csw=1

We ran the second example while looking for "related works" for the present study. We thought that the papers of interest might refer to the process at hand in alternative ways (expressed by the or term), employing verbs such as capture, or extract. Moreover, by introducing further directives, we concentrated on .edu sites and pdf files. For this query, we chose to indicate explicitly the starting day: March 7th, 2004.

```
:- prep_google(['Google snippets', 'or(capture,extract)', 'at:.edu',
               'in:english', 'file:pdf', 'since:3/7/2004'], U).

U = https://www.google.com/search?as_epq=Google+snippets&
    as_oq=capture+extract&as_sitesearch=.edu&lr=lang_en&
    as_filetype=pdf&tbs=cdr:1,cd_min:3/7/2004,cd_max: .

:- search(['Google snippets', 'or(capture,extract)', 'at:.edu',
          'in:english', 'file:pdf', 'since:3/7/2004']).
want to see keywords?[y/n] y


name    : Using syntactic and semantic relation analysis
          in question  answering.
date    : Dec 24, 2005
url     : http://www.comp.nus.edu/~kanmy/papers/2005-trec.pdf
info    : technique cannot be directly applied to extract ... dependency
          relation analysis to extract answer nuggets for ... picks
          expansion terms from Google snippets.
keywords: [dependency relation analysis, semantic relation analysis, picks
           expansion terms, answer nuggets, question answering, Google
           snippets, syntactic]
want to see?[y/n/end]:
```

The Prolog files recorded through the execution of store_sn in the two examples are shown in appendices A and B.

2.2. Implementation features

The current prototype implementation of the tool runs on a Windows platform. It is written in SWI-Prolog[4], version 5.11.14 (as updated in Jan 27th, 2011). Each Google generated-page is accessed via a special predicate https_get(U,S), where U is a url generated by prep_google (with an additional directive for controlling the access to the next pages, to be explained later) and S is a character string, consisting of the entire contents of the page in html notation.

We had first tried the http_get predicate available from the HTTP package of SWI-Prolog, but were not able to make it handle appropriately the https communication protocol, as employed by Google. Our predicate, coded in Java and functioning via the JPL interface of SWI-Prolog, utilizes a HttpURLConnection to make requests to the HTTPS server of Google. Its source code is reproduced in appendix C.

The search and the store_sn predicates, introduced in the previous section, have a control structure that enables them to access, one by one, each page obtained by the Google search. After a page is retrieved by https_get, our program extracts, also one by one, the snippets contained in the page. The total number of snippets per page is 10, this being a default that we decided to keep, but of course can be less than that in the last page.

[4] http://www.swi-prolog.org/

Due to our requirement that the snippets be dated, apparently a consequence of our adoption of the starting date directive, one way to locate the beginning of a snippet is to look for a '<h3 class="r"><a href=' substring. Of course this is guaranteed to work only with the current version of Google, and is subject to change, which also applies to all the other delimiters whereupon our extraction method is now based.

Recognizing the occasional need for adaptation, we carefully modularized our program, providing separate auxiliary predicates, such as get_sn to extract an entire snippet, and get_name, get_url, get_info, etc. for the various components, so that the places to update could be more easily located. And whenever it becomes necessary to inspect the internal structure of the pages resulting from a search, one can perform a query directly via the Google interface, click on the right side of the mouse, and select the "view source" option.

The keywords specific to each resource are currently extracted by the AlchemyAPI service.[5] We found expedient to join the name and the info fields to form a single string, that is then submitted for keyword extraction.

The control mechanism, whereby the search and store_n predicates are made to iterate across successive pages, relies on the serial number of the first snippet of the page to be accessed, which is 0 for the first page, with increments of 10 for the next ones. With variable Ipg representing the serial number, the url produced by prep_google is concatenated to '&start=', Ipg, '&num=10' before https_get is executed to fetch a page.

We found that Google currently does not indicate failure when an order to fetch more pages is issued after the last one has been processed. To avoid that this last page be treated unendingly as a new one, our mechanism stops the process as soon as the first snippet of two apparently distinct consecutive pages is found to contain the same value in the url field.

.

2.3. Limitations and extensions

A word of caution is in order. The tool must be used sparingly, since Google will block what it detects as a series of attempts by a "machine" (rather than a human agent) to access its services, whenever this goes beyond an internally established threshold in terms of number of accesses per time interval. So, any person or organization intending to make massive usage, even for strictly research purposes, of tools like ours should first contact the closest authorized Google representative. Also, especially if commercial applications are involved, a similar precaution is recommended towards the other organizations whose products (in principle offered on a free albeit limited basis) are part of the tool, in our case SWI-Prolog and the AlchemyAPI keyword extractor.

In order to establish restrictions on our own experiments, we programmed the search and the store_sn predicates to never go beyond 10 Google pages (each page with 10 snippets, as noted before). The starting date was not limited, but, if left unspecified, we take the preceding 10 years as default.

Google snippets sometimes offer, in separate positions, complementary information such as authors' names, etc., which the tool does not presently catch. Moreover some differently formatted snippets are not processed at all, particularly the commercial ads. If needed, each of these now absent elements may well be considered without difficulty in later versions, simply by finding the appropriate delimiting tags in the Google source pages and plugging them into the program so as to handle the adopted varieties of formats. Despite these lacks,

[5] http://www.alchemyapi.com/api/keyword-extraction/

we think that what we can already cover is ample enough to justify our belief in the usefulness of our approach.

One more serious limitation derives from the very notion of snippet. A snippet, contrary to an abstract or summary, is by definition a fragment with interruptions signaled by suspension marks, not a regular text with meaningful sentences fully conforming to the rules of the language grammar. Although they are usually effective as a sample, sufficient to help the user decide whether or not a resource meets the objective of the search, they can be occasionally more intriguing than informative.

Even keyword extraction may suffer from the fragmentary nature of snippets. The AlchemyAPI extractor's announcement, for instance, leaves clear how critically it depends on the ability to examine the information contents: "We employ sophisticated statistical algorithms and natural language processing technology to analyze your data, extracting keywords that can be used to index content, generate tag clouds, and more".

Fortunately such limitations can be partly counterbalanced by an extended application of the tool. Consider the first snippet obtained in our example 1:

```
name   : A Keyword-based Guide to Poirot Stories - PUC-Rio
date   : Jul 29, 2013
url    : ftp://ftp.inf.puc-rio.br/pub/docs/techreports/13_10_lima.pdf
info   : associated with the story, which include but are not limited to plot-
         summaries, narrative texts and .... kws(1, investigation, [clue:
         'character of the victim', snag: 'time of death', ... With the choices 7,
         12 for victim, 3, -11, 14 for crime, and 1, 2, 8 for.
keywords: [narrative texts, Poirot Stories, Keyword-based Guide,victim, kws, snag,
           PUC-Rio, clue, choices, story,plot-summaries, investigation, character,
           time, death, crime]
want to see?[y/n/end]:
```

If one wishes to see the continuation of the first sentence in the info field, there is the possibility to learn more by using the fragment itself in a new query, whose formulation and result is shown below. Notice that, besides finishing the sentence (and starting another one...), the result provided a few extra keywords.

```
search(['not limited to plot-summaries, narrative texts and']).
want to see keywords?[y/n] y

name     :   A Keyword-based Guide to Poirot Stories - PUC-Rio
date     :   Jul 29, 2013
url      :   ftp://ftp.inf.puc-rio.br/pub/docs/techreports/13_10_lima.pdf
info     :   but not limited to plot-summaries, narrative texts, and videos; and
             (2) keywords of different classes,
             which serve as a multi-aspect index mechanism. The system ...
keywords :  [multi-aspect index mechanism, Poirot Stories, Keyword-based Guide,
             narrative texts, different classes, PUC-Rio, keywords, plot-summaries,
             videos]
```

We experimented with another possible extension to help answering a frequent question: given two keywords K1 and K2, find in what ways the objects that they denote are related. Let, for instance, K1 = Agatha Christie and K2 = Hercule Poirot. A perhaps naive but still useful way to start attacking the problem is to find, in a series of snippets, what appears between K1 and K2 in the info fields resulting from an appropriately formulated query. Here Google greatly facilitates the task by providing a "wildcard" notation: K1 * K2. Conveniently, the entire matching word sequences are represented in bold font – in terms of html tags, between <em> and </em> (or alternatively between <b> and </b>).

To execute the task, we wrote predicate find_rel, coded with the basic predicates supplied by the tool. It considers sequences beginning with <em>K1 that contain K2 and terminate at the nearest occurrence of </em>. This criterion may look a bit unnatural. Would it not be simpler to look for sequences beginning with <em>K1 and ending with K2</em>? The problem is that K2 may figure with some sort of suffix – as for instance in the genitive form Hercule Poirot's, in which case the sequence would be wrongly rejected. On the other hand, wrong acceptance would happen if the delimiters <em>K1 and </em> were chosen without testing for the occurrence of K2, because isolated occurrences of K1 are also represented in bold font.

Applying this method, the tool found a number of word sequences connecting the famous writer with the no less famous little Belgian, taken from pages published by the British newspaper *Daily Mail*. The calling command, the generated query expression, and the resulting sequences, taken from a single Google page, are displayed below.

```
:- prep_google(['Agatha Christie * Hercule Poirot', 'at:dailymail.co.uk',
               'in:english', 'since:1/1/2013'],U),
   find_rel(U, 'Agatha Christie','Hercule Poirot',Rs), nl, nl,
   forall(member(M,Rs),(write(M),nl,nl)).

U = https://www.google.com/search?as_epq=Agatha+Christie+*+Hercule+Poirot&
    as_sitesearch=dailymail.co.uk&lr=lang_en&tbs=cdr:1,cd_min:1/1/2013,cd_max:,

Agatha Christie created Hercule Poirot

Agatha Christie's detective Hercule Poirot

Agatha Christie, Belgian detective Hercule Poirot

Agatha Christie's charismatic detective Hercule Poirot

Agatha Christie's effete Belgian detective Hercule Poirot

Agatha Christie described Hercule Poirot's

Agatha Christie could not stand Hercule Poirot
```

The first sequence provides a straightforward answer to the question, connecting the writer and her personage by the obvious created relation. The next four sequences classify Poirot as a detective, indicate his Belgian nationality, and point out a few of his peculiar characteristics. The sixth sequence is a case where K2 has trailing characters. And it stops in midair: it does not tell us what is being described by the author – it is Poirot's 'rapid, mincing gait', as we can learn, as we did before, by submitting the interrupted sentence to the search predicate:

```
:- search(['Agatha Christie described Hercule Poirot''s']).
want to see keywords?[y/n] y

name  : Poirot actor David Suchet on how he perfected signature walk ...
date  : Nov 5, 2013
url   : http://www.capitalbay.com/news/412618-poirot-actor-david-suchet-on-how-he-
perfected-signature-walk.html
info   : Agatha Christie described Hercule Poirot's 'rapid, mincing gait' in her
novels; The 67-year-old actor used Christie's description as his inspiration; He
repeatedly ...
keywords: [Poirot actor David,Hercule Poirot,Agatha Christie,signature walk,67-
year-old actor]
want to see?[y/n/end]:
```

and the seventh sequence is by far the most remarkable, expressing the ambiguous sentiment of the creator for her 'insufferable' creature.

3. Exploiting the snippet files

3.1. Direct usage

Even with no more than the built-in features of SWI-Prolog, the user is already able to handle the sn clauses of a snippet file in useful ways. For instance, the lines below will exhibit the names of the resources whose keyword list explicitly mentions "Poirot":

```
:- findall(N,(criminal:sn([name:N,_,_,_,kws:K]),
              member(M,K),sub_string(M,_,_,_,'Poirot')),Ns),
   setof(Ni,member(Ni,Ns),Ns1),
   forall(member(Ni,Ns1),(write(Ni),nl,nl)).
```

```
A Keyword-based Guide to Poirot Stories - PUC-Rio

Agatha Christie Poirot: The Movie Collection, Set 5 (Third Girl ...

Agatha Christie's Poirot - The Definitive Collection Series 1-13 DVD ...

Agatha Christie's Poirot: The Movie Collection - Set 4 : DVD Talk ...

Amazon.co.uk: Customer Reviews: Agatha Christie's Poirot - The ...

Celebrating Films of the 1960s & 1970s - Entries from December 2013

Creator/Agatha Christie - Television Tropes & Idioms

Department of English and American Studies ^Faculty of Arts ...

Free forensics Essays and Papers - 123HelpMe.com

Hercule Poirot: Facts, Discussion Forum, and Encyclopedia Article

Murder in the Mews - Wikipedia, the free encyclopedia

Mythodea (Music For The NASA Mission: 2001 Mars Odyssey)

Nigel Bromley - Agatha Christies - cath and nigel's home page

Poirot: Series 10 Blu-ray - Blu-ray.com

Previously undiscovered Agatha Christie works published for the ...

Reconstruction 11.3 (2011): Gender and Popular Fiction, edited by ...

SOUND INSIGHTS: April 2010

See related - Hachette Children's Books

Series/Poirot - Television Tropes & Idioms

Table of contents - RUDAR

The Adventure of the Christmas Pudding - Wikipedia, the free ...

The Dulcinea Effect - Welcome to the Tropes Mirror Wiki on Wikia!
```

```
Variations on Three Bodies of Knowledge | van der Linde ...

kinds of narrators and focalisers
```

Interpretive languages, such as Prolog, allow the user to view a result and immediately employ it in other non-anticipated tasks. Noticing the reference to the story "Murder in the Mews" as part of the answer to the previous query, the user may wish to read its plot, which is most likely to be present in the resource, since its name field reveals that it is a Wikipedia page. For opening the page, the win_shell built-in predicate can be promptly applied to the contents of the respective url field:

```
:- criminal:sn([name:'Murder in the Mews - Wikipedia, the free
                encyclopedia',_,url:U,_,_]),
   win_shell(open,U).
```

Far more power is added if the snippet file is loaded together with the **LOG-SNIP** tool, since the search/store facilities of the latter can work on various combinations of the extracted keyword lists. The following lines select the keywords that occur in at least 8 of the items kept in our criminal.pl snippet file, and calls search over an and combination of these frequently occurring keywords:

```
findall(K,(criminal:sn([_,_,_,_,kws:Ks]),member(K,Ks)),Kt),
setof(E-C,
     (I,Is)^(member(E,Kt),findall(1,member(E,Kt),Is),length(Is,C)),Kt1),
findall(K,(member(K-C,Kt1),C >= 8),Kx),
L =.. [and|Kx], term_to_atom(L,L1), search([L1]).
```

noting that, using the three keywords thus obtained, the final call above to the search predicate is executed over the argument ['and(Agatha Christie, Poirot stories, victim)']). The first two resulting snippets are:

```
name  : A Time-Lapse Detective: 25 Years of Agatha Christie's "Poirot" |
date  : Nov 25, 2013
url   : https://lareviewofbooks.org/essay/a-time-lapse-detective-25-years-of-
agatha-christies-poirot
info  : Agatha Christie, after Shakespeare and the authors of the Bible, ranks as
the third ..... This final season has been striking for the attention paid to the
victim's bodies, and their ..... Suchet had long set his sights on filming all of
the Poirot stories.
keywords: [Agatha Christie,Poirot stories,Time-Lapse Detective,final
season,sights,victim,Shakespeare,attention,bodies,authors,Bible]
want to see?[y/n/end]:

name  : Poirot - Agatha Christie - The official information and community site
date  : Sep 3, 2013
url   : http://www.agathachristie.com/christies-work/detectives/poirot/1
info  : In fact, Agatha Christie later wrote that Poirot's introduction to
detective fiction was not at all ... It took David Suchet almost exactly 25 years
to film 70 Poirot stories ...
keywords: [Agatha Christie,Poirot stories,David Suchet,official
information,community site,fact,introduction,fiction]
want to see?[y/n/end]: end
```

Sometimes the keyword list extracted from the snippet coming from a given resource, and made available in the respective sn clause, may be judged insufficient. In such cases, having access to the entire text or, at the very least, to some sort of summary might be a better option. Since the url field of the same sn clause gives the address of the resource, one should be able to look into its contents for this purpose, but the type of the file or its protection against unauthorized access may prevent that. For technical papers, fortunately, several organizations, such as ACM, publicize html index pages for certain papers, wherein their abstracts are displayed.

With this in mind, suppose we try again for related works (as we did in section 2.1), this time using the search list ['Google snippets', 'or(capture,extract)', 'at:dl.acm.org','since:3/7/2004'], and create a one-snippet file to serve as example. Appendix D shows how to penetrate into the located ACM page to fetch an indexed paper's abstract and then apply to it the AlchemyAPI keyword extractor, thereby obtaining a considerably larger keyword list.

3.2. Translating the snippet files to apply the KW-GPS tool

Finer-grained exploitation is achievable if the keywords extracted from the snippets are first divided into classes that are chosen in view of an application. As a consequence of this admittedly ad-hoc approach, different classifications can be envisaged depending on the user's preferences and current objectives. Although recognizing the advantages of borrowing from widely adopted ontologies, our option, for the moment at least, has been for arbitrary choices of classes, letting taxonomy take the form of folksonomy [Damme].

To provide an input to our **KW-GPS** tool, which was built to support multiple-class keywords, the predicate below translates a snippet file produced by **LOG-SNIP** into a new file with the required organization.

transf(F1, F2, Lc) - where F1 is a snippet file from which file F2 is obtained by asking the user to decide: **1.** whether or not a snippet taken from F1 should be recorded in F2 and, if the answer is positive, **2.** in which of the classes in the list Lc each keyword of the recorded snippet should be allocated.

The user has always the option to simply disregard a keyword. Certain criteria, such as TF-IDF [Wu], can help evaluating the relevance of a keyword in a given domain, but in the current version of the transf predicate the decision to retain or drop a keyword is left to the user's discretion. In case of doubt about its meaning, typing a question mark allows to consult DBpedia (preferred for words beginning with an uppercase letter) or WordNet (for lowercase).

In the example shown next, each snippet in file criminal.pl is submitted to the user's examination in order to create (or totally replace, if created before) the my_crimes.pl file, ready to be later handled by **KW-GPS**. Three keyword classes have been indicated: `personage`, `criminal`, `general`, the first for names of personages, the second for terms formally or informally related to crimes, and the third for other terms that may seem of enough interest. A few steps of the ensuing dialogue are illustrated next; the first two snippets are left out, whilst the third is accepted and the user is asked to classify its keywords – three are left out and one – Mrs. Clayton – is retained, duly classified as a personage.

```
:- transf('Criminal', 'My Crimes', [personage, criminal, general]).

item: 'A Keyword-based Guide to Poirot Stories - PUC-Rio'
want to use it?[y/n/end]:

item: 'Murder in the Mews - Wikipedia, the free encyclopedia'
want to use it?[y/n/end]:

item: 'The Adventure of the Christmas Pudding - Wikipedia, the free ...'
want to use it?[y/n/end]: y
```

```
*** please choose and classify keywords for this item ***

1:personage
2:criminal
3:general
4:none
choose for 'series Agatha Christie':

1:personage
2:criminal
3:general
4:none
choose for 'Poirot stories':

1:personage
2:criminal
3:general
4:none
choose for 'Trefusis shows':

1:personage
2:criminal
3:general
4:none
choose for 'Mrs Clayton': 1
```

The entire file, my_crimes.pl, generated through the dialogue is reproduced in appendix E. The clauses resulting from the first accepted snippet are shown below. The lib clause contains all elements taken from the snippet except the keywords, which are kept in separate kws clauses corresponding to the three classes.

```
lib( 1, 'The Adventure of the Christmas Pudding - Wikipedia, the free ...',
       [date: 'Nov 24, 2007 ',
        url: 'http://en.wikipedia.org/wiki/The_Adventure_of_the_
              Christmas_Pudding',
        info: 'Contents. 1 Plot summaries .... He is able to start investigating
               the case when a mutual friend recommends him to Mrs Clayton. ...
               Trefusis shows Poirot the scene of the crime and the detective is
               puzzled as to why there is a ..... All five of the Poirot stories
               were adapted to television as part of the series Agatha
               Christie\'s Poirot.']).

kws(1,personage,['Mrs Clayton','Poirot']).
kws(1,criminal,[detective,scene,crime]).
kws(1,general,['Christmas Pudding']).
```

Since the features of the **KW-GPS** tool were fully described in a previous document[6], we shall only present a few simple examples of their application over the my_crimes.pl clauses. The interactive execution of rank(Items) causes the resources to be evaluated according to the current user's preferences, indicated by choosing among the keywords of the three classes. As a result, the resources are listed in decreasing order of total number of hits.

```
:- rank(Items).

1:Ariadne
2:Captain Hastings
3:Dr. Stavros Constantine
```

[6] ftp://ftp.inf.puc-rio.br/pub/docs/techreports/13_10_lima.pdf

```
4:Hastings
5:Holmes
6:Miss Springer
7:Mrs Clayton
8:Mrs. McGinty
9:Oliver
10:Poirot
11:Superintendent Spence
12:Tommy
choose for personage: 7, 1, 11, 4

1:Asshole Victim
2:Criminal Investigation
3:Poirot interviews
4:Technology
5:better mysteries
6:crime
7:crime scene
8:criminal activity
9:criminal gang
10:death
11:detective
12:detective investigating murder
13:doubt
14:frequent victim
15:head photographs
16:identification
17:multiple murderer
18:piracy
19:respective investigations
20:scene
21:victim
22:victims
23:witnesses
choose for criminal: 6, 10, 11, 20, 23

1:Cat
2:Christmas Pudding
3:Dulcinea Effect
4:Elephants
5:Free forensics Essays
6:Pigeons
7:SOUND INSIGHTS
8:Savior Complex
9:Shining Armour
10:career
11:crew
12:moneylender
13:packing
14:passengers
choose for general: 2, 3, 4, 6, 14

The Adventure of the Christmas Pudding - Wikipedia, the free ... with 5 hits
SOUND INSIGHTS: April 2010 with 3 hits
```

Notice that, wherever a keyword selected by the user is absent it is simply not counted. The tool allows other options: a plus or a minus sign written before the number designating a keyword serves to indicate, respectively, that its presence is mandatory or that items containing it are to be rejected.

By applying another predicate of **KW-GPS** to a given a resource, the other resources are evaluated for their similarity with it, letting similarity be measured in terms of the number of keywords in common. The next example finds the resource that is most similar, in this sense, to the first resource in the my_crimes.pl file.

```
:- lib(1,N1,_), similar(1,[I2|_]), lib(I2,N2,_).

SOUND INSIGHTS: April 2010 with 2 hits
```

The third example shows a selection using the keywords of the three classes explicitly nominated by the user. Misspellings are automatically corrected if the Levenshtein distance [Navarro] is no greater than 2.

```
:- select([['Hastings', 'Oliver'],
           [piracy, death, victims],
           [Elephants', 'Pigeons', passengers]],
          Items).

Amazon.co.uk: Customer Reviews: Agatha Christie's Poirot - The ... with 4 hits
Series/Poirot - Television Tropes & Idioms with 3 hits
Nigel Bromley - Agatha Christies - cath and nigel's home page with 1 hits

Items = [4,3,2].
```

The **KW-GPS** tool also makes provision for keywords represented as terms, such as motive:'financial gain', motive:'moral reasons', etc. Taking advantage of this structured term format, the select predicate allows more elaborate queries by including terms with variables in the keyword lists, such as motive:M, which will match those two motive instances. Moreover, from the same snippet file any number of **KW-GPS** files can be generated with different keyword classes, customized to the taste of individual users and adequate to meet the purposes of specific applications.

4. Concluding remarks

Our proposed approach uses a list of keywords, which supposedly characterize the domain of current interest, to conduct a Google search during which the snippets of the located resources are collected. Our **LOG-SNIP** tool gives the option to store in the format of frame-structured Prolog clauses the snippets that seem more promising.

Such clauses can then be submitted to further analysis. In particular, the keyword lists extracted from the snippets of each resource serve, after being divided into separate classes, to perform aspect-oriented comparisons among the resources, employing our previously developed **KW-GPS** tool.

Achieving better versions of **LOG-SNIP** depends on a more detailed knowledge of Google, including its full set of search parameters, methods and algorithms, and the internal html structure of the resulting pages.

Until now the tool runs exclusively in the standard SWI-Prolog environment. To fit it for practical usage in realistically ample scale, appropriate menu-driven user interfaces need to be developed for each specific application.

References

[Bratko] I. Bratko. *Prolog Programming for Artificial Intelligence*. Pearson Education Canada, 2011.

[Christie] A. Christie. *Hercule Poirot - the Complete Short Stories*. Harper, 2008.

[Damme] V. C. Damme, M. Hepp, K. Siorpaces. "FolksOntology: An integrated approach for turning folksonomies into ontologies". *Proc. of ESWC 2007 - Bridging the Gap between Semantic Web and Web 2.0 workshop*, 2007.

[Hessick] C.B. Hessick. "Motive Role in Criminal Punishment". *Southern California Law Review*, pp. 89-150, 2008.

[Lima] E.S. Lima, B. Feijó, S.D.J Barbosa, A.L, Furtado. "A Keyword-based Guide to Poirot Stories", *Technical Report 10/13*, Departamento de Informatica, PUC-Rio, 2013.

[Navarro] G. Navarro. "A Guided Tour to Approximate String Matching". *ACM Computing Surveys*, vol. 33 (1), pp. 31-88, 2001.

[Ribeiro-Neto] B. Ribeiro-Neto. "Web Search - Challenges and Opportunities". *Proc. of AMW*, 2012.

[Todorov] T. Todorov. *The Poetics of Prose*. Cornell University Press, 1977.

[Wu] H.C. Wu, R.W.P. Luk, K.F. Wong, K.L. Kwok. "Interpreting TF-IDF Term Weights as Making Relevance Decisions". *ACM Transactions on Information Systems*, vol. 26 (3), pp.1-13, 2008.

Appendix A

Snippets in the criminal domain

```
% Criminal

criminal:search_list('Criminal',['and(summary,victim,crime,investigation)','Poirot
                      stories','-blog','in:english']).
criminal:search_date(3/28/2014).
criminal:search_url('https://www.google.com/search?as_q=summary+victim+crime+investigation&
                     as_epq=Poirot+stories&as_eq=blog&lr=lang_en&
                     tbs=cdr:1,cd_min:3/28/2004,cd_max:').

:- dynamic criminal:sn/1 .
:- retractall(criminal:sn(_)).
:- set_prolog_flag(toplevel_print_options,[max_depth(50)]).

criminal:sn([name  : 'A Keyword-based Guide to Poirot Stories - PUC-Rio',
    date  : 'Jul 29, 2013 ',
    url   : 'ftp://ftp.inf.puc-rio.br/pub/docs/techreports/13_10_lima.pdf',
    info  : 'associated with the story, which include but are not limited to plot-summaries,
             narrative texts and .... kws(1, investigation, [clue: \'character of the
             victim\', snag: \'time of death\', ... With the choices 7, 12 for victim, 3,
             -11, 14 for crime, and 1, 2, 8 for.',
    kws   : ['narrative texts', 'Poirot Stories', 'Keyword-based Guide', victim,kws,snag,
             'PUC-Rio', clue, choices, story, 'plot-summaries', investigation, character,
              time, death, crime]]).

criminal:sn([name  : 'Murder in the Mews - Wikipedia, the free encyclopedia',
    date  : 'May 14, 2008 ',
    url   : 'http://en.wikipedia.org/wiki/Murder_in_the_Mews',
    info  : 'The victim was locked in her room and was shot through the head with an ... He
             is stunned to find out that a murder investigation is taking place and admits
             .... Wishing for a quiet holiday free from crime, Poirot goes to Rhodes during
             the low ... the four awkwardly shaped Poirot stories which make up Murder in
             the Mews that I ...',
    kws   : ['awkwardly shaped Poirot','murder investigation','quiet holiday','free
              encyclopedia','Mews',admits,victim,'Rhodes',room,head,place,crime,stories]]).

criminal:sn([name  : 'The Adventure of the Christmas Pudding - Wikipedia, the free ...',
    date  : 'Nov 24, 2007 ',
    url   : 'http://en.wikipedia.org/wiki/The_Adventure_of_the_Christmas_Pudding',
    info  : 'Contents. 1 Plot summaries .... He is able to start investigating the case when
             a mutual friend recommends him to Mrs Clayton. ... Trefusis shows Poirot the
             scene of the crime and the detective is puzzled as to why there is a ..... All
             five of the Poirot stories were adapted to television as part of the series
             Agatha Christie\'s Poirot.',
    kws   : ['series Agatha Christie', 'Poirot stories', 'Trefusis shows', 'Mrs Clayton',
             'mutual friend', 'Christmas Pudding', 'Plot summaries', detective, 'Contents',
              case, scene, crime, television, 'Poirot', 'Adventure']]).

criminal:sn([name  : 'Crimes of Conscience: Morality and Justice in Doyle ... -
                      Crimeculture',
    date  : 'Jul 18, 2008 ',
    url   : 'http://www.crimeculture.com/Contents/Articles-Summer07/Crimes_Conscience.html',
    info  : 'The police, of course, want to capture the criminal and see legal justice ...
             are portrayed as bumbling conventionalists, victims of their own orthodoxy, and
             it is up ...',
    kws   : ['legal justice', orthodoxy, 'Conscience', 'Morality', 'Doyle', 'Crimes',
              victims, police, course, conventionalists]]).

criminal:sn([name  : 'Detective Fiction (Bookshelf) - Gutenberg',
    date  : 'Oct 23, 2011 ',
    url   : 'http://www.gutenberg.org/wiki/Detective_Fiction_(Bookshelf)',
    info  : 'Detective Fiction is a branch of crime fiction that centers upon the
             investigation of a crime, usually murder, by a detective, either professional
             or amateur.',
    kws   : ['Detective Fiction', 'crime fiction', 'Bookshelf', murder, branch,
              investigation, 'Gutenberg']]).

criminal:sn([name  : 'Variations on Three Bodies of Knowledge | van der Linde ...',
    date  : 'Apr 6, 2004 ',
    url   : 'http://journals.hil.unb.ca/index.php/IFR/article/view/7743/8800',
    info  : 'Knowledge that the investigator has prior to the investigation includes
             specialized ... from previous investigations, containing information on crimes,
             criminal types, ... or clarify the profile of the victim; or to open up a line
             of investigation based on a ..... In the Poirot stories, the main problem is
             fairly clear-cut, but the detective\'s ...',
    kws   : ['van der Linde', 'Poirot stories', 'previous investigations', 'criminal types',
             'main problem', investigator, 'Knowledge', victim, crimes, detective,
             'Variations', 'Bodies', information, profile]]).

criminal:sn([name  : 'Nigel Bromley - Agatha Christies - cath and nigel\'s home page',
    date  : 'Aug 14, 2004 ',
    url   : 'http://www.caffnib.co.uk/agathaps.shtml',
    info  : 'She and Tommy help to uncover a criminal gang and also unmask a multiple
             murderer, ... it must be one of the passengers or crew; the victim was a
             moneylender and possibly a .... Superintendent Spence thinks he didn\'t do it
             and asks Poirot to investigate. ... Hero: Poirot; Summary: A collection of four
             short Poirot stories.',
    kws   : ['short Poirot stories', 'Nigel Bromley', 'Agatha Christies', 'Superintendent
             Spence', 'multiple murderer', 'criminal gang', 'home page', cath, 'Tommy',
             passengers, victim, 'Hero', crew, moneylender, 'Summary']]).

criminal:sn([name  : 'Table of contents - RUDAR',
    date  : 'Jun 13, 2007 ',
    url   : 'http://rudar.ruc.dk/bitstream/1800/2617/1/SYNOPSIS%20Suspense%20and%20
             Surprise%20in%20Five%20in%20Agatha%20Christie%20Stories%20til%20
             biblioteket.pdf',
    info  : 'crime novel to hold its reader in suspense all the way through and still
             surprise. .... The Poirot-stories are typical of the British whodunit detective
             story concerned with who did it and ... assassin did not know what the left
             hand was doing to the victim. ... The titles of the chapters resemblance a
             summary of the investigation.',
    kws   : ['British whodunit detective', 'chapters resemblance', 'crime novel', 'left
              hand', suspense, assassin, victim, reader, way, 'Poirot-stories', story,
              titles, summary, investigation, 'Table', contents]]).

criminal:sn([name  : 'Series/Poirot - Television Tropes & Idioms',
    date  : 'Jun 9, 2013 ',
    url   : 'http://tvtropes.org/pmwiki/pmwiki.php/Series/Poirot',
    info  : 'Asshole Victim: The Miss Springer in Cat Among the Pigeons; Mrs. .... few of
             the original Poirot stories, but feature in the majority of the pre-season IX
             episodes anyway. ... Oliver\'s and Poirot\'s respective investigations in
             Elephants Can Remember, ... Pie in the Sky . CrimeAndPunishmentSeries/Detective
             Drama . Red Riding ...',
    kws   : ['original Poirot stories', 'Asshole Victim', 'pre-season IX episodes', 'Miss
              Springer', 'Television Tropes', 'respective investigations',
             'CrimeAndPunishmentSeries/Detective Drama', 'Pigeons', majority, 'Mrs',
             'Elephants', 'Pie', 'Oliver', 'Cat', feature, 'Sky']]).

criminal:sn([name  : 'Creator/Agatha Christie - Television Tropes & Idioms',
    date  : 'Jan 20, 2014 ',
    url   : 'http://tvtropes.org/pmwiki/pmwiki.php/Creator/AgathaChristie',
    info  : 'one of the murder victims was guilty, (And Then There Were None), (Curtain)
             .... the Asshole Victim, b) the killer or c) if not the killer, then a weak
             criminal type anyways. .... of two lovers who need to be brought together
             during their investigations. .... Also subverted in at least two Poirot
             stories, where a smashed watch is found ...',
    kws   : ['weak criminal type', 'Asshole Victim', 'murder victims', 'Poirot stories',
            'Television Tropes', 'Creator/Agatha Christie', killer, 'Curtain', lovers,
             watch, anyways]]).

criminal:sn([name  : 'Poirot: Series 10 Blu-ray - Blu-ray.com',
    date  : 'Dec 13, 2013 ',
    url   : 'http://www.blu-ray.com/movies/Poirot-Series-10-Blu-ray/81674/',
    info  : 'Comic book (4133) Coming of age (1417) Crime (16626) ... Overview . Blu-ray
             review . Screenshots, (40), User reviews . Region coding . News . Forum ...',
    kws   : ['Blu-ray review', 'Comic book', 'Region coding', 'Blu-ray.com', 'Poirot', age,
             'Crime', 'Overview', 'Screenshots', 'User', 'News', 'Forum']]).

criminal:sn([name  : 'Agatha Christie\'s Poirot - The Definitive Collection Series 1-13 DVD
                    ...',
    date  : 'Jan 6, 2014 ',
    url   : 'http://www.amazon.co.uk/Agatha-Christies-Poirot-Definitive-
             Collection/dp/B00EQ30DDQ',
    info  : 'The Poirot stories are really five star, but I have to remove a star for the
             ... death but still having the time to investigate some of the better mysteries
             of his career. ... and Hastings among his many victims are the treats offered
             by the final season. ... What a difference it is to watch a good crime drama
             without TV advertisements.',
    kws   : ['Poirot stories', 'Definitive Collection Series', 'good crime drama', 'Agatha
              Christie', 'better mysteries', 'final season','TV advertisements', 'Hastings',
              treats, difference, star, victims]]).

criminal:sn([name  : 'Amazon.co.uk: Customer Reviews: Agatha Christie\'s Poirot - The ...',
    date  : 'Jan 6, 2014 ',
    url   : 'http://www.amazon.co.uk/product-reviews/B00EQ30DDQ',
    info  : 'The Poirot stories are really five star, but I have to remove a star for the
             packing and that piracy warning. ... death but still having the time to
             investigate some of the better mysteries of his career. ... and Hastings among
             his many victims are the treats offered by the final season. .... are justly
             and swiftly punished for their crimes.',
    kws   : ['Poirot stories', 'Agatha Christie', 'better mysteries', 'final season',
             'Customer Reviews', 'Amazon.co.uk', piracy, star, 'Hastings', treats, packing,
              victims, death, time, career]]).

criminal:sn([name  : 'Department of English and American Studies ^Faculty of Arts ...',
    date  : 'May 22, 2011 ',
    url   : 'https://is.muni.cz/th/74475/ff_b/B.A._Thesis.txt',
    info  : 'Last reference is to Agatha Christie and brief introduction to her location.
             .... Another debate was about the type of crime that should be investigated.
             ..... and that the victim is neither likeable nor admirable but a foolish,
             domineering snob. ...... alone, he often has some assistant -- in Christie\'s
             early Poirot stories Captain Arthur ...',
    kws   : ['Poirot stories Captain', 'Agatha Christie', 'American Studies Faculty',
             'domineering snob', 'brief introduction', victim, debate, Arthur', reference,
              location, type, crime, 'Department', 'English']]).

criminal:sn([name  : 'Agatha Christie Poirot: The Movie Collection, Set 5 (Third Girl ...',
    date  : 'Jul 23, 2010 ',
    url   : 'http://www.dvdtalk.com/reviews/42812/agatha-christies-poirot-the-movie-
             collection-set-5/',
    info  : 'Finally, recent Poirot stories have been firmly entrenched in the late-1930s
             instead of ... Investigating further - with Ariadne\'s assistance - Poirot
             interviews Norma\'s .... After going over the crime scene with Dr. Stavros
             Constantine (Samuel West), ... Should not the victims and grieving survivors
             therefore be entitled to mete out ...',
    kws   : ['recent Poirot stories','Agatha Christie Poirot','Dr. Stavros
              Constantine','Poirot interviews','Samuel West','crime scene','Movie
              Collection','Ariadne']]).

criminal:sn([name  : 'Agatha Christie\'s Poirot: The Movie Collection - Set 4 : DVD Talk
             ...',
    date  : 'Jun 30, 2009 ',
    url   : 'http://www.dvdtalk.com/reviews/37664/agatha-christies-poirot-the-movie-
             collection-set-4/',
    info  : 'Intriguingly, one of the major differences seems to be that while Christie\'s
             Poirot stories ... Bulstrode engages Poirot to assist in the investigation, and
             individually ... The murders are vividly staged and, while hardly deserving of
             their fate, the victims are not exactly tragic figures. ... A Brief History of
             Time: Criterion Collection ...',
    kws   : ['Poirot stories', 'exactly tragic figures', 'Agatha Christie', 'major
              differences', 'Criterion Collection', 'Brief History','DVD Talk','Movie
              Collection', 'Bulstrode', murders, fate, victims, investigation, 'Time']]).

criminal:sn([name  : 'Reconstruction 11.3 (2011): Gender and Popular Fiction, edited by
                    ...',
    date  : 'Sep 28, 2011 ',
    url   : 'http://reconstruction.eserver.org/113/Walker.shtml',
    info  : 'Carpan provides a decade-by-decade overview of trends in girls\' series, but in
             such ... his youthful identity as a teen safely investigating the world within
             the confines of his ... such as cults and séances, associated with Golden Age
             crime 2) ethnic and ... Curran received permission to reprint two uncollected
             Poirot stories and ...',
    kws   : ['uncollected Poirot stories', 'Golden Age crime', 'youthful identity', 'decade-
              by-decade overview', 'girls series', 'Popular Fiction', confines, 'Curran',
              cults, permission, trends, 'Reconstruction', 'Carpan', world]]).

criminal:sn([name  : 'Free forensics Essays and Papers - 123HelpMe.com',
    date  : 'Mar 8, 2011 ',
    url   : 'http://www.123helpme.com/search.asp?text=forensics',
    info  : '... will help create a primary overview before anything has been touched or
             processed. ... How Technology Has Impacted Criminal Investigation - . ....
             Holmes by Sir Arthur Conan Doyle, Hercule Poirot stories by Agatha Christie
             which are .... from head photographs to be released to the media for
             identification of the victim.',
    kws   : ['Hercule Poirot stories', 'Arthur Conan Doyle', 'Free forensics Essays',
             'Agatha Christie', 'primary overview', 'Criminal Investigation', 'head
              photographs', 'Holmes', 'Sir', 'Technology', media, identification, victim]]).

criminal:sn([name  : 'Previously undiscovered Agatha Christie works published for the ...',
    date  : 'Aug 28, 2011 ',
    url   : 'http://www.dailymail.co.uk/news/article-2031070/Previously-undiscovered-Agatha-
             Christie-works-published-time-audio-books.html',
    info  : 'The previously unpublished Poirot stories of The Capture of Cerberus and The
             ... and to mark what would have been the Queen of Crime\'s 121st birthday. ...
             But a chance meeting with an old acquaintance leads to an investigation ......
             Co-ordination: Rescue workers look for victims in the mudslide near Oso,
             Washington ...',
    kws   : ['previously unpublished Poirot', 'Agatha Christie', 'old acquaintance', 'chance
              meeting', 'Rescue workers', 'Cerberus', 'Oso', 'Co-ordination', 'Capture',
              birthday, victims, stories, 'Queen', 'Crime', investigation, mudslide]]).

criminal:sn([name  : 'See related - Hachette Children\'s Books',
    date  : 'Nov 26, 2013 ',
    url   : 'https://www.hachettechildrens.co.uk/Search.page?SearchRelated=Deadly%20
             Tales,%20Roy%20Apps,%20Ollie%20Cuthbertson,%20EDGE,%20don\'t%20look%20behind
             %20you%20and%20the%20babsitter,%20the%20bloody%20hook%20and%20vanishing
             %20hitchhiker,%20the%20party%20animal%20and%20don\'t%20look%20under%20the
             %20bed,%20the%20hangover%20and%20dead%20man%20drinking,%20love%20bites%20and
             %20it%20crawled%20from%20the%20dark,%20raijin%20and%20woman%20in%20the%20
             mirror&SearchTitle=Deadly%20Tales',
    info  : '[8] The children\'s author Anne Fine presented an overview of the concerns
             about ..... began with the adaptation of several Hercule Poirot stories for
             ITV\'s popular Agatha ... derided science-fiction shows, Crime Traveller (1997)
             for BBC One and The ... While investigating Fire and World\'s workshop, James
             is suspected of ...',
    kws   : ['Hercule Poirot stories', 'author Anne Fine', 'popular Agatha', 'science-
              fiction shows', 'Crime Traveller', 'Hachette Children', 'ITV', adaptation,
              concerns, overview, 'BBC', 'World', workshop, 'James']]).

criminal:sn([name  : 'kinds of narrators and focalisers',
    date  : 'Sep 22, 2004 ',
    url   : 'http://ieas.unideb.hu/admin/file_6355.doc',
    info  : 'The victims of sexist thought and sexual discrimination are usually women (for
             instance, ... This also implies that anthropologists investigate not only the
             so-called ..... a scene in which the detective gathers together all the
             possible suspects of the crime, ... establishes an intertextual link with
             Agatha Christie\'s Poirot stories.',
    kws   : ['sexist thought', 'Agatha Christie', 'detective gathers', 'intertextual link',
             'Poirot stories', 'sexual discrimination', 'possible suspects', narrators,
              instance, anthropologists, 'so-called']]).

criminal:sn([name  : 'SOUND INSIGHTS: April 2010',
    date  : 'Apr 29, 2010 ',
    url   : 'http://dougpayne.blogspot.com/2010_04_01_archive.html',
    info  : 'But even Agatha Christie had the “retired” detective investigating murder and
             mayhem for ... The edits are brief and unimportant to the plot and no doubt
             made to allow for ... suspected and witnesses even place her at the scene of
             the crime. ...... provides references to such previous Poirot stories as Mrs.
             McGinty\'s Dead ...',
    kws   : ['detective investigating murder', 'previous Poirot stories', 'Agatha Christie',
             'Mrs. McGinty', 'SOUND INSIGHTS', mayhem, edits, doubt, witnesses, plot, cene,
              crime, references]]).

criminal:sn([name  : 'Celebrating Films of the 1960s & 1970s - Entries from December 2013',
    date  : 'Dec 31, 2013 ',
    url   : 'http://www.cinemaretro.com/index.php?/archives/2013/12.html',
    info  : 'The Detective is blatantly breaking the law by setting up a crime and forcing
             ..... In summary, Grindhouse Releasing has outdone itself with this
             presentation of a very ... internationally renowned detective, Hercule Poirot,
             investigating a murder on ... of Christie\'s Poirot stories, including “Murder
             on the Orient Express” in 2010.',
    kws   : ['internationally renowned detective', 'Hercule Poirot', 'Poirot stories',
             'Grindhouse Releasing', 'Orient Express', murder, 'Christie', '1960s']]).

criminal:sn([name  : 'The Dulcinea Effect - Welcome to the Tropes Mirror Wiki on Wikia!',
    date  : 'Dec 24, 2012 ',
    url   : 'http://tropes.wikia.com/wiki/The_Dulcinea_Effect',
    info  : 'The Knight in Shining Armour is a frequent victim of the effect. .... Doumeki -
             even though he did not want help - before investigating his other options. ...
             a Savior Complex towards any innocent getting involved with criminal activity.
             ..... Captain Hastings, from the Poirot stories by Agatha Christie, suffers
             from this a great deal.',
    kws   : ['Tropes Mirror Wiki', 'Dulcinea Effect', 'Shining Armour', 'frequent victim',
             'Agatha Christie', 'Poirot stories', 'Captain Hastings', 'Savior Complex',
             'criminal activity', 'great deal', 'Wikia', 'Knight', help, options]]).

criminal:sn([name  : 'Hercule Poirot: Facts, Discussion Forum, and Encyclopedia Article',
    date  : 'Sep 2, 2008 ',
    url   : 'http://www.absoluteastronomy.com/topics/Hercule_Poirot',
    info  : 'Overview. Hercule Poirot is a fictional Belgian detective created by Agatha ...
             A more obvious influence on the early Poirot stories is that of Arthur Conan
             Doyle ..... of the crime scene, but by enquiring either into the nature of the
             victim or the ..... Poirot travelled all over Europe and the Middle East
             investigating crimes and ...',
    kws   : ['Overview. Hercule Poirot', 'early Poirot stories', 'fictional Belgian
              detective', 'Arthur Conan Doyle', 'East investigating crimes', 'obvious
              influence', 'crime scene', 'Discussion Forum', 'Encyclopedia Article',
             'Agatha', victim, nature, 'Europe', 'Middle']]).

criminal:sn([name  : 'Mythodea (Music For The NASA Mission: 2001 Mars Odyssey)',
    date  : 'May 16, 2011 ',
    url   : 'http://supershopsite.com/Movies/vangelis-mythodea-music-for-tje-nasa-mission-
             2001.htm',
    info  : 'Index Of All Afatha Christie Hercule Poirot Stories Agatha Christie Biography
             David .... Machinery Infused With Human Nerves, Robocop Transformed Crime-
             ridden ..... Teens Make The Fatal Mistake Of Dumping Their Victim\'s Body Into
             The Sea. ... In The Tradition Of Such Films As "brief Encounter" And "lost In
             Translation".',
    kws   : ['Christie Hercule Poirot', 'Fatal Mistake', 'Human Nerves', 'Mars Odyssey',
             'Biography David', 'brief Encounter', 'NASA Mission', 'Robocop', 'Victim',
             'Teens', 'Index', 'Afatha', 'Stories', 'Machinery', 'Tradition', 'Films',
             'Body', 'Sea', 'Translation']]).
```

Appendix B

Snippets of related work

```
snippets:search_list('Snippets',['Google snippets', 'or(capture,extract)', 'at:.edu',
                                 'in:english', 'file:pdf', 'since:3/7/2004']).
snippets:search_date(3/7/2014).
snippets:search_url('https://www.google.ca/search?as_epq=Google+snippets&as_oq=capture+extra
                     ct&as_sitesearch=.edu&lr=lang_en&as_filetype=pdf&
                     tbs=cdr:1,cd_min:3/7/2004,cd_max:').

:- dynamic snippets:sn/1 .
:- retractall(snippets:sn(_)).
:- set_prolog_flag(toplevel_print_options,[max_depth(50)]).

snippets:sn([name  : 'D07-1004',
    date  : 'Jun 16, 2007 ',
    url   : 'http://acl.ldc.upenn.edu/D/D07/D07-1004.pdf',
    info  : '(http://www.google.com/apis), we extract n(= 8) different candidates from the
             top m(= 30) Google snippets. The Google snippets containing the same candidates
             ...',
    kws   : ['Google snippets','different candidates','D07-1004']]).

snippets:sn([name  : 'Mining Translations of OOV Terms from the Web through Cross ...',
    date  : 'Aug 15, 2005 ',
    url   : 'http://www.cs.cmu.edu/~fhuang/publications/SIGIR05_WebMining.pdf',
    info  : 'glish pages containing the Chinese OOV term and extract the translations from
             the ... sent to Google, snippets containing the query and possibly its English
             ...',
    kws   : ['Chinese OOV term', 'glish pages', 'Mining Translations', 'OOV Terms',
              snippets, query, 'Google', 'English', 'Web']]).

snippets:sn([name  : 'Discovering Semantic Biomedical Relations utilizing the Web',
    date  : 'Apr 5, 2008 ',
    url   : 'http://home.cc.gatech.edu/ssahay/uploads/3/final%20tkdd%20version.pdf',
    info  : 'It is very difficult to automatically extract relations from Web pages
             expressed in ..... high coverage (that is, they are able to return many correct
             Google snippets).',
    kws   : ['correct Google snippets','Semantic Biomedical Relations','high coverage','Web
              pages']]).

snippets:sn([name  : 'Creating a Dead Poets Society: Extracting a Social ... -
             ResearchGate',
    date  : 'Mar 21, 2008 ',
    url   : 'http://semarch.linguistics.fas.nyu.edu/Archive/TEwNDRlZ/geleijnse-korst-
             ISWC2007.pdf',
    info  : 'We present a simple method to extract information from search engine snippets.
             Although the ..... Google snippets found with the queried patterns. Using the
              ...',
    kws   : ['search engine snippets', 'Dead Poets Society', 'Google snippets', 'simple
              method']]).

snippets:sn([name  : 'The 4th Web as Corpus - LREC Conferences',
    date  : 'Mar 4, 2008 ',
    url   : 'http://www.cs.brandeis.edu/~marc/misc/proceedings/lrec-
             2008/workshops/W19_Proceedings.pdf',
    info  : 'We want the Demon, you see, to extract from the dance of atoms only information
             that is genuine, like ..... with Google snippets: in this case, since we had
             to.',
    kws   : ['Google snippets', 'LREC Conferences', 'atoms only information', 'Demon',
             'Corpus', dance ,case, 'to', 'Web']]).

snippets:sn([name  : 'Using syntactic and semantic relation analysis in question
             answering.',
    date  : 'Dec 24, 2005 ',
    url   : 'http://www.comp.nus.edu/~kanmy/papers/2005-trec.pdf',
    info  : 'technique cannot be directly applied to extract ... dependency relation
             analysis to extract answer nuggets for ... picks expansion terms from Google
             snippets.',
    kws   : ['dependency relation analysis', 'semantic relation analysis', 'picks expansion
              terms', 'answer nuggets', 'question answering', 'Google snippets',
              syntactic]]).

snippets:sn([name  : 'The effect of bias on an automatically-built word sense corpus',
    date  : 'Apr 13, 2004 ',
    url   : 'https://www.comp.nus.edu/~rpnlpir/proceedings/lrec-2004/pdf/648.pdf',
    info  : 'Google snippets for each monosemous word in WordNet. 1.7. Then, for each ...
             for each sense, and we extract the snippets as returned by the search engine.',
    kws   : ['word sense corpus', 'Google snippets', 'monosemous word', 'search
              engine',bias,effect]]).

snippets:sn([name  : 'Constructing Task-Specific Taxonomies for Document Collection ...',
    date  : 'Jul 11, 2012 ',
    url   : 'http://www.cs.georgetown.edu/~huiyang/publication/emnlp2012.pdf',
    info  : 'A general scheme to capture user inputs in tax- ... step is to extract the
             concepts and the second is to organize the concepts ... the top 10 Google
             snippets. (3) We ...',
    kws   : ['Task-Specific Taxonomies', 'user inputs', 'Google snippets', 'general scheme',
             'Document Collection', concepts,step]]).

snippets:sn([name  : 'Enriching short text representation in microblog for clustering',
    date  : 'Nov 7, 2011 ',
    url   : 'http://www.public.asu.edu/~jtang20/publication/short%20message.pdf',
    info  : 'that such a translation may not be able to capture accurate term meanings. ...
             [15] cluster short texts (i.e., Google snippets) by first extracting the
             important ...',
    kws   : ['accurate term meanings', 'cluster short texts', 'short text representation',
             'Google snippets', microblog]]).

snippets:sn([name  : 'The effect of bias on an automatically built word sense corpus - UPC',
    date  : 'Jan 21, 2005 ',
    url   : 'http://www.lsi.upc.edu/~nlp/meaning/documentation/3rdYear/WP6.12a.pdf',
    info  : 'In order to build this corpus3, we have acquired 1000 Google snippets for each
             ... Google5 with the monosemous relatives for each sense, and we extract the.',
    kws   : ['word sense corpus','Google snippets','monosemous relatives']]).

snippets:sn([name  : 'Semantic Term Matching in Axiomatic Approaches to Information ...',
    date  : 'Aug 11, 2006 ',
    url   : 'http://times.cs.uiuc.edu/czhai/pub/sigir06-semantic.pdf',
    info  : 'STMC3 intends to capture the following intuition: Sup- pose we ..... from the
             Google snippets. We set ... trieval constraints are defined to capture
             intuitions on se-.',
    kws   : ['Semantic Term Matching', 'following intuition', 'Google snippets', 'Axiomatic
              Approaches', 'trieval constraints', intuitions]]).

snippets:sn([name  : 'Mining User Relations from Online Discussions using Sentiment ...',
    date  : 'Apr 23, 2013 ',
    url   : 'http://www.mysmu.edu/phdis2010/minghui.qiu.2010/papers/13-NAACL-PMF.pdf',
    info  : 'We also extract opin- ions of users towards other users and .... tries and
             Google snippets may be considered. We will study this problem in our future
             work.',
    kws   : ['opinions', 'Mining User Relations', 'Google snippets', 'future work', 'Online
              Discussions','Sentiment',tries,users,problem]]).
```

Appendix C

Fetching web pages using the https protocol

```
/* prolog predicate */
/* employing the Prolog/Java interface */

:- use_module(library(jpl)).

https_get(U, C) :-
   jpl_new('https_get', [], P),
   jpl_call(P, geturl, [U], C).


/* the Java source program */

import java.net.HttpURLConnection;
import java.net.URL;
import java.io.BufferedReader;
import java.io.InputStreamReader;
import java.io.UnsupportedEncodingException;

public class https_get
{

   public static String geturl(String url) throws Exception
   {
        URL website = new URL(url);
        String result = "";
        HttpURLConnection connection = (HttpURLConnection)
website.openConnection();
        connection.addRequestProperty("User-Agent", "Mozilla/5.0 (Windows NT 6.1;
Win64; x64; rv:25.0) Gecko/20100101 Firefox/25.0");
        BufferedReader in = new BufferedReader(new
InputStreamReader(connection.getInputStream()));

        StringBuilder response = new StringBuilder();
        String inputLine;
        while ((inputLine = in.readLine()) != null)
            response.append(inputLine);
        in.close();

        try
        {
          result = new String(String.valueOf(response).getBytes(), "UTF-8");
        }catch(UnsupportedEncodingException uee){
            uee.printStackTrace();
        }

        return result;
    }
}
```

Appendix D

Extracting keywords from an abstract

```
:- store_sn('One Snippet',['Google snippets', 'or(capture,extract)',
                'at:dl.acm.org','since:3/7/2004']).

name  : Web-based pattern learning for named entity translation in ...
date  : 'Mar 2, 2009 '
url   : http://dl.acm.org/citation.cfm?id=1464526.1465266
info  : These patterns can be used to extract K-C, K-E and E-C pairs from Google snippets. We
found KCIR performance using this hybrid configuration over five times ...
kws   : [E-C pairs,Google snippets,hybrid configuration,entity translation,Web-based
pattern,KCIR performance,K-C,patterns,K-E]
want to store?[y/n/end]: y


:- consult(one_snippet).

:- one_snippet:sn([_,_,url:Ux,_,_]),
   atom_concat(Ux,'&preflayout=flat',Ux1),
   https_get(Ux1,T1),
   (sub_string(T1,I,_,_,'<p>'),!,
    sub_string(T1,J,_,_,'</p>'),
    D = 3;
    sub_string(T1,I,_,_,'<div style="display:inline">'),
    sub_string(T1,J,_,_,'</div>'),
    D = 28),
      I < J,
    I1 is I + D,
    L is J - I1,
    sub_string(T1,I1,L,_,X1), !,
   html_to_txt(X1,X),
   text_kws(X,[Kws]).

X = Named entity (NE) translation plays an important role in many applications, such as
information retrieval and machine translation. In this paper, we focus on translating NEs
from Korean to Chinese in order to improve Korean-Chinese cross-language information
retrieval (KCIR). The ideographic nature of Chinese makes NE translation difficult because
one syllable may map to several Chinese characters. We propose a hybrid NE translation
system. First, we integrate two online databases to extend the coverage of our bilingual
dictionaries. We use Wikipedia as a translation tool based on the inter-language links
between the Korean edition and the Chinese or English editions. We also use Naver.com's
people search engine to find a query name's Chinese or English translation. The second
component of our system is able to learn Korean-Chinese (K-C), Korean-English (K-E), and
English-Chinese (E-C) translation patterns from the web. These patterns can be used to
extract K-C, K-E and E-C pairs from Google snippets. We found KCIR performance using this
hybrid configuration over five times better than that a dictionary-based configuration using
only Naver people search. Mean average precision was as high as 0.3385 and recall reached
0.7578. Our method can handle Chinese, Japanese, Korean, and non-CJK NE translation and
improve performance of KCIR substantially.

Kws = [NE translation, hybrid NE translation, makes NE translation, information retrieval,
cross-language information retrieval, non-CJK NE translation, Mean average precision,
English-Chinese E-C, E-C pairs, ideographic nature, machine translation, inter-language
links, important role, dictionary-based configuration, Korean-English K-E, Chinese
characters, Naver people, bilingual dictionaries, translation tool, online databases, hybrid
configuration, English translation, Korean edition, English editions, Google snippets,
translation patterns, KCIR performance, K-C].
```

Appendix E

File generated by the transf predicate to serve as input to the KW-GPS tool

```
/* domain My Crimes */

:- set_prolog_flag(toplevel_print_options,[max_depth(50)]).
:- style_check([-singleton,-discontiguous]).

kw_classes([personage,criminal,general]).
thresholds(St,[P,C,G],Any).
keyword_lists(St,[P,C,G]).

% LIBRARY

lib( 1, 'The Adventure of the Christmas Pudding - Wikipedia, the free ...',
       [date: 'Nov 24, 2007 ',
        url: 'http://en.wikipedia.org/wiki/The_Adventure_of_the_
              Christmas_Pudding',
        info: 'Contents. 1 Plot summaries .... He is able to start investigating
               the case when a mutual friend recommends him to Mrs Clayton. ...
               Trefusis shows Poirot the scene of the crime and the detective is
               puzzled as to why there is a ..... All five of the Poirot stories
               were adapted to television as part of the series Agatha
               Christie\'s Poirot.']).

lib( 2, 'Nigel Bromley - Agatha Christies - cath and nigel\'s home page',
       [date: 'Aug 14, 2004 ',
        url: 'http://www.caffnib.co.uk/agathaps.shtml',
        info: 'She and Tommy help to uncover a criminal gang and also unmask a
               multiple murderer, ... it must be one of the passengers or crew;
               the victim was a moneylender and possibly a .... Superintendent
               Spence thinks he didn\'t do it and asks Poirot to investigate. ...
               Hero: Poirot; Summary: A collection of four short Poirot
               stories.']).

lib( 3, 'Series/Poirot - Television Tropes & Idioms',
       [date: 'Jun 9, 2013 ',
        url: 'http://tvtropes.org/pmwiki/pmwiki.php/Series/Poirot',
        info: 'Asshole Victim: The Miss Springer in Cat Among the Pigeons; Mrs.
               .... few of the original Poirot stories, but feature in the
               majority of the pre-season IX episodes anyway. ... Oliver\'s and
               Poirot\'s respective investigations in Elephants Can Remember, ...
               Pie in the Sky . CrimeAndPunishmentSeries/Detective Drama . Red
               Riding ...']).

lib( 4, 'Amazon.co.uk: Customer Reviews: Agatha Christie\'s Poirot - The ...',
       [date: 'Jan 6, 2014 ',
        url: 'http://www.amazon.co.uk/product-reviews/B00EQ30DDQ',
        info: 'The Poirot stories are really five star, but I have to remove a
               star for the packing and that piracy warning. ... death but still
               having the time to investigate some of the better mysteries of his
               career. ... and Hastings among his many victims are the treats
               offered by the final season. .... are justly and swiftly punished
               for their crimes.']).

lib( 5, 'Agatha Christie Poirot: The Movie Collection, Set 5 (Third Girl ...',
       [date: 'Jul 23, 2010 ',
        url: 'http://www.dvdtalk.com/reviews/42812/agatha-christies-poirot-the-
              movie-collection-set-5/',
        info: 'Finally, recent Poirot stories have been firmly entrenched in the
               late-1930s instead of ... Investigating further - with Ariadne\'s
               assistance - Poirot interviews Norma\'s .... After going over the
               crime scene with Dr. Stavros Constantine (Samuel West), ... Should
               not the victims and grieving survivors therefore be entitled to
               mete out ...']).

lib( 6, 'Free forensics Essays and Papers - 123HelpMe.com',
       [date: 'Mar 8, 2011 ',
        url: 'http://www.123helpme.com/search.asp?text=forensics',
        info: '... will help create a primary overview before anything has been
               touched or processed. ... How Technology Has Impacted Criminal
               Investigation - . .... Holmes by Sir Arthur Conan Doyle, Hercule
               Poirot stories by Agatha Christie which are .... from head
               photographs to be released to the media for identification of the
               victim.']).

lib( 7, 'SOUND INSIGHTS: April 2010',
       [date: 'Apr 29, 2010 ',
        url: 'http://dougpayne.blogspot.com/2010_04_01_archive.html',
        info: 'But even Agatha Christie had the “retired” detective investigating
               murder and mayhem for ... The edits are brief and unimportant to
               the plot and no doubt made to allow for ... suspected and
               witnesses even place her at the scene of the crime. ......
               provides references to such previous Poirot stories as Mrs.
               McGinty\'s Dead ...']).

lib( 8, 'The Dulcinea Effect - Welcome to the Tropes Mirror Wiki on Wikia!',
       [date: 'Dec 24, 2012 ',
        url: 'http://tropes.wikia.com/wiki/The_Dulcinea_Effect',
        info: 'The Knight in Shining Armour is a frequent victim of the effect.
               .... Doumeki - even though he did not want help - before
               investigating his other options. ... a Savior Complex towards any
               innocent getting involved with criminal activity. ..... Captain
               Hastings, from the Poirot stories by Agatha Christie, suffers from
               this a great deal.']).


% KEYWORDS

kws(1,personage,['Trefusis shows','Mrs Clayton','Poirot']).
kws(1,criminal,[detective,scene,crime]).
kws(1,general,['Christmas Pudding']).


kws(2,personage,['Superintendent Spence','Tommy']).
kws(2,criminal,['multiple murderer','criminal gang',victim]).
kws(2,general,[passengers,crew,moneylender]).


kws(3,personage,['Miss Springer','Oliver']).
kws(3,criminal,['Asshole Victim','respective investigations']).
kws(3,general,['Pigeons','Elephants','Cat']).


kws(4,personage,['Hastings']).
kws(4,criminal,['better mysteries',piracy,victims,death]).
kws(4,general,[packing,career]).


kws(5,personage,['Dr. Stavros Constantine','Ariadne']).
kws(5,criminal,['Poirot interviews','crime scene']).
kws(5,general,[]).
kws(6,personage,['Holmes']).
kws(6,criminal,['Criminal Investigation','head
photographs','Technology',identification,victim]).
kws(6,general,['Free forensics Essays']).


kws(7,personage,['Mrs. McGinty']).
kws(7,criminal,['detective investigating murder',doubt,witnesses,scene,crime]).
kws(7,general,['SOUND INSIGHTS']).


kws(8,personage,['Captain Hastings']).
kws(8,criminal,['frequent victim','criminal activity']).
kws(8,general,['Dulcinea Effect','Shining Armour','Savior Complex']).
```