



Automatic categorisation for IPSV users

Contents

1	Introduction to IPSV	2
2	What is automatic categorisation and why have it?	2
3	How does automatic categorisation work?	3
3.1	Rule-based categorisation	3
3.2	Statistical methods	3
3.3	Neural networks.....	4
3.4	Linguistic analysis.....	4
3.5	Which approach is best?	4
4	How well does automatic categorisation work?	4
5	Features to look out for, in a categorisation package	4
6	Evaluation	5
7	References and links to IPSV documentation.....	5

Revision history of this Guide

Revision date	Previous revision date	Summary of changes
2006-03-31		First issue
2006-08-29	2006-03-31	Updated to reflect the release of e-GMS Version 3.1, and remove an Appendix

Metadata

Coverage.spatial	UK
Creator	Stella Dextre Clarke, Consultant
Date.issued	2006-03-31
Date.modified	2006-08-29
Description	Advice on automating the meta-tagging process with IPSV (Integrated Public Sector Vocabulary)
Format	Text
Language	eng
Publisher	Porism Limited, London, SW9 8BJ info@porism.com
Rights.copyright	Crown copyright. This document may be used and copied without payment or licence for research, private study or for internal circulation within an organisation. This is subject to the material being reproduced accurately and not used in a misleading context. The source and copyright status must be acknowledged in any copy of whole or part of the document. Any other proposed use of the document requires a copyright licence, which is available from www.opsi.gov.uk
Status	Version 1.1. For publication
Subject	Controlled vocabularies; Metadata; Automatic indexing
Title	Automatic categorisation for IPSV users
Type	Instructional

1 Introduction to IPSV

The Integrated Public Sector Vocabulary (IPSV) is a structured list of terms for the Subject metadata of public sector resources. Use of IPSV is part of the compliance requirements for the e-Government Metadata Standard (e-GMS). It is also mandated by the Department for Communities and Local Government (DCLG) (formerly the Office of the Deputy Prime Minister) for use by local authorities in England.

The purpose of using IPSV terms is to improve retrieval of information resources from websites, intranets, and any other collection. To make this possible, each web page or document needs to be categorised. In other words, at least one relevant preferred term from IPSV needs to be included in the Subject metadata for the page.

IPSV is governed by a Board with representatives from the e-Government Unit (e-GU), Department for Communities and Local Government, esd-toolkit, the executive authorities in Wales, Scotland and Northern Ireland, the National Archives (TNA) and the Office of Public Sector Information (OPSI)¹, local government and the trade association Intellect. IPSV is published online, with a navigable/searchable tree structure and a range of downloadable formats, on the esd-toolkit's standards site, at <http://www.esd.org.uk/standards/ipsv/>.

2 What is automatic categorisation and why have it?

The process of entering Subject metadata for a document is often called "meta-tagging" or "indexing" or (less commonly) "categorisation". Traditionally it is done by a person, who reads the document to decide what it is about before choosing the right Subject metadata. When the process

¹ It is anticipated that OPSI will merge with TNA in October 2006

is automated it is usually called “automatic categorisation”, and sometimes shortened to “auto-categorisation”. This Guide complements the [Guide to Meta-tagging with the IPSV](#), which sets out the basic technique of indexing, by adding guidance on automation of the process.

The purpose of automation is to save time, money and hassle. Traditional indexing may require a trained person to spend several minutes on each document that enters the system. If there are only 100 documents in the system, and each takes an average of just five minutes to index, that adds up to 500 minutes of work, or 8.4 hours. Furthermore, indexers who are not trained or not motivated can make a very poor job of it. Potentially, automation can save time, reduce errors and improve consistency.

However, the savings will only be realised if

- the price is right
- you implement an efficient workflow that does not incur higher costs.
- the quality of the results is good enough

You should satisfy yourself on all these points before you implement.

This guidance note assumes you will be applying the auto-categorisation to IPSV. But remember you may be able to use the same methodology/software to meta-tag with other controlled vocabularies such as the Local Government Classification Scheme (LGCS), or the Local Government Navigation List (LGNL), or your own specialised vocabulary, at little extra cost. For each different application you may need to vary the workflow to maximise efficiency – think carefully about this before you implement.

3 How does automatic categorisation work?

Four main approaches are used by categorisation software:

- Rule-based
- Statistical
- Neural networks
- Linguistic

3.1 Rule-based categorisation

Using the first approach, a rule is established for each category in the vocabulary. The rule is used to test the full text of the document. For example, the category “motorcycles” might have a rule: “Look for occurrences of ‘motorcycle’ or ‘motorcycles’ or ‘motorbike’ or ‘motorbikes’ or ‘motor cycle’ or ‘motor cycles’ or ‘motor-bike’ or ‘motor-bikes’, and the more occurrences there are, the higher the probability that the document should be categorised “motorcycles”.

Depending on the capabilities of the software, the rule can be more sophisticated, e.g. it could look for ‘AIDS’ but ignore ‘aids’, or it could look for the word ‘band’ only if it is close to another word such as ‘music’ or ‘jazz’ or ‘brass’. Sometimes the rule can give more weight to the word if it is found in the title or first paragraph, than if it appears in the body of the text.

An important feature of a good package is that the rules give a weighted result, so that documents that match the criteria can be ranked with the most relevant ones first, and/or the categories that get a low weighting can be ignored.

The rule is not usually expressed as an ordinary sentence, but in a compressed syntax proprietary to the software. However, it should be easy enough for a trained person to read, and edit the rule if it is giving poor results.

3.2 Statistical methods

With statistical or probabilistic categorisation, each category has attached to it an algorithm that is not so easy for an editor to adjust. Typically the application is “trained” to recognise each category, by presenting to it a number of documents that have already been judged relevant. The software analyses the documents and works out what features distinguish the relevant documents from the rest of the collection. Again, it takes into account occurrences and co-occurrences of particular words and perhaps punctuation or location of words in the document. Several different statistical approaches may be found among rival products.

When the algorithm cannot be easily edited, the usual method of correcting unsatisfactory performance is to change the 'training set' of documents, and retrain the software with new documents that give better results.

3.3 Neural networks

This approach involves statistical analysis too, as well as an architecture of interconnected circuits, modelled on the human brain, to detect patterns in the data. Like most of the statistical approaches, the software needs to learn to recognise each category by analysing a training set of correctly categorised documents.

3.4 Linguistic analysis

Linguistic analysis enables the software to distinguish nouns from verbs or adjectives, so for example it can tell the difference between a 'painting' as a work of art, and the activity of 'painting'. It can also distinguish isolated words from the same words in meaningful phrases, for example a phrase such as 'cost benefit analysis' or 'green belt' has much more significance than the individual words that make it up.

Linguistic analysis is usually combined with statistical data to develop an algorithm for each category, which cannot be adjusted directly. Again the usual method of correcting poor categorisation is to change the 'training set' for the category concerned.

3.5 Which approach is best?

Each of these approaches has its strengths and weaknesses. Some commercial products offer a combination of the basic methods, and it is hard to compare their merits. It is safest to evaluate software offerings with a collection of your own documents. Do not rely on the vendor's demonstration using an orderly sample such as the Reuter's news feed.

4 How well does automatic categorisation work?

Some people argue that a machine is incapable of understanding what a piece of text is really about, and so the job should be done by humans. Others point out that human indexers tend to be inconsistent, and indeed tests have shown that it is hard to achieve consistency higher than 60%, between the results of two different indexers. But machines sometimes get things consistently wrong!

Another common argument is that the cost of human indexing is unaffordable; better to have machine indexing than none at all. Furthermore, it can be hard to find people with the training and motivation to do the job properly.

There is no black-and-white answer to these arguments. The best thing is to weigh up the cost, feasibility and quality of results that can be obtained in your situation.

You may conclude that the best solution is a mix. For example, you might use automatic categorisation to get your legacy collection very quickly indexed, but use human indexing for new documents coming in. Or you might use automatic categorisation to "suggest" categories, but a human operator checks the results and can override any that seem anomalous.

5 Features to look out for, in a categorisation package

- Can the vendor supply the package with IPSV already built in?
- Has the package already been trained (or the rule-base built) to categorise with IPSV? Does it make use of IPSV features such as the non-preferred terms and scope notes?
- How easy is it to update the package when IPSV is updated?
- Does the vendor take responsibility for implementing IPSV updates promptly?
- How much adaptation will be needed to adjust to your own situation? For example, your documents may be significantly different from the norm, and the rule-bases/training may need to be adjusted. You may have local terms to add in to the rule-bases. How conveniently can you add your adaptations, and maintain them each time IPSV is updated?

- Is it easy to import any other taxonomies you are using, such as your own specialised thesaurus?
- How much instruction will your staff need to be able to build rule-bases and/or adjust the training sets? Does the vendor provide learning support?
- Does the package provide good diagnostics while the package is being trained or customised? (e.g. It should help you work out why a particular rule is giving anomalous results.)
- Does the package allow a human indexer or quality controller to override false categorisations?
- How well does the package or module integrate with other functions in the meta-tagging workflow?
- How big does the training set have to be, for each new category?
- Do you know other purchasers who are prepared to share their experiences with you?
- Before you commit to purchase, can you see a trial run with your own collection?

All these features can have a big effect on your costs, and the quality of your results. Don't just accept the vendor's assurances without testing/demonstration.

A starter list of automatic categorisation products may be found in the esd-toolkit Guide, *Software for automatic categorisation with IPSV*.

6 Evaluation

It is important to "try before you buy", using your own test sample. Some vendors will let you run a test at no cost. Preferably, you should select a representative sample of documents or pages and get them categorised carefully by someone who is trained in indexing. (You could try the [Society of Indexers](#) if you don't have in-house expertise. Look for someone who knows your subject area and has a background in this type of indexing rather than back-of-the-book indexing.) Compare their results with those of the products you have short-listed. Try to avoid investing huge development effort in a new product before its effectiveness can be properly evaluated.

7 References and links to IPSV documentation

Name of document	Comment
IPSV FAQs	The easiest place to start
Guide to Meta-tagging with the IPSV	Advice for webmasters and authors of electronic resources when entering metadata
Implementing IPSV to your own advantage	Hints to help you get the most out of IPSV
IPSV Guidance Notes	Especially useful for local authority users
Design/selection criteria for software used to handle controlled vocabularies	Helps with choosing software for any part of the implementation
IPSV Editorial Policy	Includes discussion of the issues leading to development of the policies now in force
IPSV Maintenance Guide	Useful for the IPSV editor and for developers of other category lists, thesauri, etc.
Which IPSV? A guide to the versions and formats available	Full description of the options: online display, downloadable files, full or abridged, machine-readable or for human eyes
e-Government Metadata Standard (e-GMS)	Full details of all the metadata elements needed for interoperability in the public sector

Note: Comments on this Guide should be posted on the IPSV [discussion forum](#)