Dynamic Change in Classification Systems: A Preliminary Model for Instance-based Matching

Fredric C. Gey
UC Data Archive & Technical Assistance
University of California, Berkeley
gey@berkeley.edu

ABSTRACT

This paper presents a preliminary mathematical model for classification systems which change over time. Capturing this dynamic change is useful in instance-based matching from older versions of a classification system to newer versions. Such mappings have practical implications for the development of historical economic time series when the classification of economic activity has undergone change.

1 Introduction

Classification systems have been utilized to organize and summarize information both semantically and statistically. For example the Medical Subject Headings (MeSH) system has been used for decades to index the medical literature, while the International Patent Classification system[1] (IPC) indexes patents by the technological areas to which they pertain. Economic activity has also been subjected to classifications such as the North American Industrial Classification System (NAICS)[2] which is used to classify companies and statistically summarize (aggregate) their economic activity (sales, employment, profits) into readily identifiable and semantically meaningful categories (e.g. NAICS code 336111 -- Automobile Manufacturing). While some would claim [Soergel 1999] that ontologies are a re-invention of classification, classification systems may be viewed as a special case of ontologies where the relationships may or may not be explicitly defined.

In order to represent the real world, classification systems must adapt to changing conditions. For example until the 1983, AIDS did not exist as a classification within MeSH. Indeed for 2007, National Library of Medicine announced[3] that over 615 MeSH classifications were being modified (494 descriptors were added, 99 descriptor terms were replaced with more up-to-date terminology, and 22 descriptors were deleted). This represents a 2 ½ percent change to the 24,357 descriptors in MeSH at the time. For economic classifications, until 1997 all USA economic activity was summarized by the Standard Industrial Classification (SIC) system. Beginning in 1997, SIC was replaced by NAICS, a system developed jointly with agencies from Canada, the USA and Mexico in order to achieve comparable reporting of economic activity in North America. This represented a massive change in economic reporting for all three countries. It is these dynamic changes which this paper attempts to model.

[1] http://www.wipo.int/classifications/ipc/en/
[2] http://www.census.gov/epcd/www/naics.html
[3] http://www.nlm.nih.gov/pubs/techbull/nd06/nd06_mesh.html

Instance based mapping utilizes the characteristics of instances classified by particular classification systems to create the mapping between the classification systems. According to Isaac and others [Isaac et al 2007] “there are surprisingly few systematic studies of instance-based ontology mappings, i.e. the construction of links between concepts based on the co-occurrence of instances.” This author has been working on creating a syntax to capture instance-based mappings for classification systems within the ISO 11179 standard for metadata registries [Gey 2008].

2 Static models for instance based mapping between classification systems

We assume that there exist two classification schemes C and D such that

- For each classification concept, cj € C and dk € D there exist a set I (database) of instances which are associated with (classified by) cj and dk
- The entire universe of instances are classified by both C and D**.**

2.2 Simple Static Model

The mapping between two concepts cj € C and dk € D is created by using the *count of instances* held in common between the two classifications

- Example 1: a bibliographic database of books is indexed by two different classification schemes: Library of Congress classification and Dewey Decimal System.
- Example 2: all businesses in the United States are assigned unique SIC and NAICS codes

2.2.1 Simple Static Mapping

- the *degree* to which cj maps to dk is the percent of instances indexed by cj which are also indexed by dk
- the degree to which cj maps from dk is the percent of instances indexed by dk which are also indexed by cj

2.2.2 Associative Static Mapping

In the simple static mapping of the last section, we directly utilize the count (or percent) of instances in common. However, if one classification is utilized more frequently than another, the overlap disproportionately influences the degree of mapping. This tendency can be compensated for by using statistical association measures such as Yates Chi Square [Yates 1934] or Dunning’s log likelihood [Dunning 1994]. As a precursor to computing these measures between each classification concept, we need to create a 2-

way contingency table of counts of instances between each classification pair cj € C and dk € D, as in the following table:

Classifi-cation	d_k	$\neg d_k$
c_j	n_1	n_2
$\neg c_j$	n_3	n_4

Table 1: Contingency Table Summarizing Classification Instance Counts

Where n_l is the number of instances classified by both cj and dk, n_2 is the number of instances classified by cj but not classified by dk, n_3 is the number of instances classified by dk, but not classified by cj and n_4 is the number of instances classified by neither of the two classifications.

2.2.3 Instance-Based Mapping: The Weighted Static Model

The weighted static model extends the basic static model to weight each instance count by some numeric weight which has meaning. For example, in mapping instances of businesses indexed by two economic classifications (such as SIC and NAICS above), we could weight each business by its number of employees or by its total sales. More formally, as before,

- there exist two classification schemes ***C* and *D*** such that
  - For each classification **cj € *C*** and **dk € *D*** there exists a set ***I*** (database) of instances which are associated with (classified by) **cj** and **dk**
  - The entire universe of instances are classified by ***C*** and ***D.***
  - Each instance **i € *I*** has a numeric weight **w_i**
- then for **Weighted Static Mapping**
  - the ***degree*** to which **<u>cj maps to dk</u>** is the ratio of weighted sum of instances indexed by cj over the weighted sum instances indexed by dk, i.e. $\sum$ **w_i € cj /**$\sum$ **w_i € dk** expressed as a percent
  - the ***degree*** to which **<u>cj maps from dk</u>** is the ratio of weighted sum of instances indexed by dk which are also indexed by cj (also expressed as percent)

2.2.4 The Weighted Associative Model

The weighted associative model extends the weighted static model to allow for computation of association measures as above in the static model. In this case for each classification pair **cj € *C*** and **dk € *D,*** we create a weighted contingency table where instead of counts, sums of weights are placed in the four cells of the contingency table:

Classification	$\mathbf{d_k}$	$\neg \mathbf{d_k}$
$\mathbf{c_j}$	$\sum_{j,k€C\cap D} w_{j,k}$	$\sum_{j,k€C\cap\neg D} w_{j,k}$
$\neg\mathbf{c_j}$	$\sum_{j,k€\neg C\cap D} w_{j,k}$	$\sum_{j,k€\neg C\cap\neg D} w_{j,k}$

Table 2: Contingency Table Summarizing Classification Weights

3 Dynamic models for instance based classification mappings

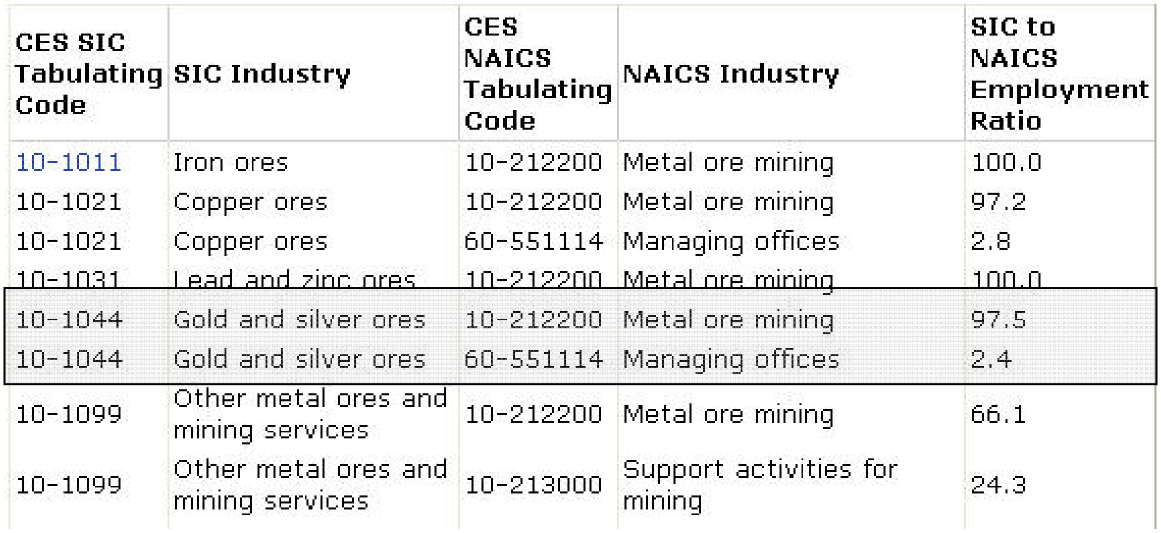
Thus far we have assumed that both the classifications in a classification system and the instances which they classify are static for all time. This is hardly the case, as we noted in the introduction for MeSH classification. Instances being classified are also dynamic. In an expanding universe such as web pages, new urls are constantly created and, in some cases, older pages disappear. For economic statistics, new business establishments are created by entrepreneurs and many of these businesses fail to stand the test of time. Indeed, it has been documented that for the time period 1993-2002, approximately 70 percent of all new businesses in the USA folded within the ten year time frame [Shane 2008]. Clearly the real-world picture for instance-based matching between classification schemes must strive to capture these changes.

3.2 A Simple Dynamic Weighted Model

The simplest dynamic model assumes that the classification systems and their instances remain unchanged, but that the weights of instances indexed by these classifications vary over time. Thus, for example, each year a business establishment will have a different number of employees and different sales, profit and debt figures, but the systems for classifying businesses and summarizing their activity is static. Only in this case, the weights (i.e. sales, count of employees, etc) must are dynamically time-varying. More formally, as before,

- There exist two classification schemes ***C* and *D*** such that
  - For each classification **cj €** ***C*** and **dk €** ***D*** there exists a set ***I*** (database) of instances which are associated with (classified by) **cj** and **dk**
  - The entire universe of instances are classified by ***C*** and ***D.***
  - *Each instance* **i € I** *has a time-varying numeric weight* $\mathbf{w_i(t)}$
- **Simple dynamic weighted mapping**
  - the *degree* to which cj maps to dk ***at time t*** is the ratio of weighted sum of instances indexed by cj over the weighted sum of instances indexed by dk expressed as a percent
  - the *degree* to which cj maps from dk ***at time t*** is the ratio of weighted sum of instances indexed by dk which are also indexed by cj (also expressed as percent**)**
- **Approximating for a single point in time,** $\mathbf{t_n}$**, we assume** $\mathbf{w_i(t) \approx w_i(t_n)}$

An example from the USA Bureau of Labor Statistics (BLS) may be helpful in illustrating this case. In 2001, BLS tabulated its "Quarterly Census of Employment and Wages Program" using both Standard Industrial Classification Codes and North American Industrial Classification Codes. From this tabulation BLS prepared a detailed table of matching ratios between each SIC code and each NAICS code using number of employees as the weight factor for each industry (business establishment). The following figure is a fragment of this table for the Mining sector:



CES SIC Tabulating Code	SIC Industry	CES NAICS Tabulating Code	NAICS Industry	SIC to NAICS Employment Ratio
10-1011	Iron ores	10-212200	Metal ore mining	100.0
10-1021	Copper ores	10-212200	Metal ore mining	97.2
10-1021	Copper ores	60-551114	Managing offices	2.8
10-1031	Lead and zinc ores	10-212200	Metal ore mining	100.0
10-1044	Gold and silver ores	10-212200	Metal ore mining	97.5
10-1044	Gold and silver ores	60-551114	Managing offices	2.4
10-1099	Other metal ores and mining services	10-212200	Metal ore mining	66.1
10-1099	Other metal ores and mining services	10-213000	Support activities for mining	24.3

Figure 1: SIC to NAICS mapping example for Mining Industries

The highlighted section shows that SIC classification code 1044, Gold and silver ores maps to two NAICS Codes 212200 (Metal ore mining) and 55114 (Managing offices) to degrees 97.5 percent and 2.4 percent respectively. Thus from the above, we could say:

- **for the BLS example t_n = 2001**

The full table may be found at: http://www.bls.gov/ces/sic4tonaics.htm.

3.3 An Extended Dynamic Weighted Model

An extended dynamic model would assume that the classification systems remain unchanged, but their instances change over time. Thus, for example, new business establishments are created and establishments which fail are no longer accounted for, but the systems for classifying businesses and summarizing their activity is static. Thus we may assume for a single classification scheme:

- For each classification cj € C there exists a set I_{cj} (database) of instances which are associated with (classified by) cj
- The universe of instances classified by cj ***varies with time***
- Thus $I_{cj}(t_n)$ associated with cj € C and $I_{cj}(t_{n+1})$ are not necessarily identical
- Example from economy: 70 percent of all new businesses fail within 10 years. Yet these are often replaced (in an expanding economy) by new businesses.
- Each instance i € I_{cj} has a time-varying numeric weight $w_i(t)$

While, in principle, this seems to be a more complex case, the formulae for allocation of degree of match remain the same as in the simple dynamic weighted model, because the weights are still summed over the universe $\mathbf{I_{cj}(t_n)}$ at each time $\mathbf{t_n}$

3.4 *The Full Dynamic Weighted Model*

In the full dynamic weighted model we are faced with how to capture not only the characteristics of instance universes change but also of time-varying changes in the classification systems themselves, i.e. new classifications (e.g. internet service providers, cell phones) come into existence and old ones (buggy whips) disappear. Thus:

- There exist two classification schemes C and D which ***are time varying, i.e. they are more accurately characterized as C(t) and D(t)***
- For each classification cj € C(t) and dk € D(t) there exists sets $I_{cj}(t)$ and $I_{dk}(t)$ (database) of instances which are associated with (classified by) cj and dk
- The universes of instances classified by C and D also vary with time, i.e.
  $I_{cj}(t_n)$ € C and $I_{dk}(t_n)$ € D not necessarily identical with $I_{cj}(t_{n+1})$ or $I_{dk}(t_{n+1})$
  Each instance i € I has a time-varying numeric weight $w_i(t)$

There are numerous complexities in trying to do instance-based matching across time variant classification systems. For example, discontinuity -- if a classification which existed at time $\mathbf{t_n}$ and does not exist at time $\mathbf{t_{n+1}}$, should the instances at time $\mathbf{t_n}$ be allocated to a different classification at time $\mathbf{t_{n+1}}$? The author is still conceptualizing what instance-based matching should be for this general case.

4 Summary

The goal of this paper has been to describe change in classification systems and describe mathematical models for instance-based matching between such systems. Motivation for systematizing such matching can be found in economics when attempting to create consistent historical economic time series when the underlying classifications of economic activity have undergone change.

5 Acknowledgments

This research was partially supported by USA National Science Foundation Grant 0637122 “Ecoinformatics Project: Semantics Management and Semantics Services.”

6 References

[Dunning 1994] T Dunning, Accurate Methods for the Statistics of Surprise and Coincidence, *Computational Linguistics*, v19, no 1, 1994, pp 61-74.

[Gey 2008] F Gey, Syntaxes for Mapping Between Classification Systems,” *International Journal of Metadata, Semantics and Ontologies,* forthcoming, Winter 2009.

[Isaac et al 2007] A Isaac and others, An empirical study of instance-based ontology matching, in The Semantic Web, 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007 + ASWC 2007, Busan, Korea, November 11-15, 2007, pp 253-266.

[Shane 2008] S Shane, *Illusions of Entrepreneurship: The Costly Myths that Entrepreneurs, Investors, and Policy Makers Live By*. Yale University Press, January 28, 2008.

[Soergel 1999] D Soergel, The Rise of Ontologies or the Reinvention of Classification. JASIS 50 (12): 1119-1120 (1999).

[Yates 1934] Yates, F. Contingency tables involving small numbers and the χ^2 test, *Journal of the Royal Statistical Society* (Supplement) **1**: 217-235.