Interpolation and Extrapolation in Conceptual Spaces: A Case Study in the Music Domain

Steven Schockaert[1] and Henri Prade[2]

[1] Department of Applied Mathematics and Computer Science, Ghent University, Belgium, steven.schockaert@ugent.be
[2] Institut de Recherche en Informatique de Toulouse (IRIT), Université Paul Sabatier, Toulouse, France, prade@irit.fr

Abstract. In most knowledge representation settings, atomic properties correspond to natural language labels. Although these labels are usually taken to be primitive, automating some forms of commonsense inference requires background knowledge on the cognitive meaning of these labels. We consider two such forms of commonsense reasoning, which we refer to as interpolative and extrapolative reasoning. In both cases, rule-based knowledge is augmented with knowledge about the geometric representation of labels in a conceptual space. Specifically, to support interpolative reasoning, we need to know which labels are conceptually between which other labels, considering that intermediary conditions tend to lead to intermediary conclusions. Extrapolative reasoning is based on information about the direction of change that is needed when replacing one label by another, taking the view that parallel changes in the conditions of rules tend to lead to parallel changes in the conclusions. In this paper, we propose a practical method to acquire such knowledge about the conceptual spaces representation of labels. We illustrate the method in the domain of music genres, starting from meta-data that was obtained from the music recommendation website last.fm.

1 Introduction

Symbolic approaches to knowledge representation typically start from a finite set of labels, which are used to refer to properties (or concepts) from some domain of interest. For example, in a first-order setting, labels may refer to predicates such as *blue*, *small* or *expensive*. Typically, these labels can be organized in classes, such that two labels of the same class talk about the same attributes.

Example 1. Let us consider the following classes:

$$Housing = \{castle, villa, rowhouse, apartment, studio\}$$
$$Location = \{centre, outskirts, suburbs, country\}$$
$$Comfort = \{exclusive, luxurious, very\text{-}comfortable, comfortable, basic\}$$

Logical formulas may then be used to encode how labels from different classes are related to each other. Consider for instance the following knowledge base,

containing observations about the comfort level of some housing options:

$$villa(x) \wedge suburbs(x) \rightarrow luxurious(x) \quad (1)$$

$$apartment(x) \wedge suburbs(x) \rightarrow basic(x) \quad (2)$$

$$apartment(x) \wedge centre(x) \rightarrow very\text{-}comfortable(x) \quad (3)$$

Clearly, the knowledge base in the previous example is incomplete, in the sense that the comfort level of some configurations cannot be deduced from it. For instance, we have no information at all about the comfort level of an apartment in the outskirts. Intuitively, given that an apartment in the suburbs is *basic* and one in the centre is *very-comfortable*, we may think that an apartment in the outskirts would normally be *basic*, *comfortable* or *very-comfortable*. Such a commonsense inference is based on the idea of interpolation of knowledge. In particular, it relies on the assumption that intermediary conditions lead to intermediary conclusions. Clearly, this requires that a notion of betweenness can meaningfully be defined for labels of the same class. As another example, consider the comfort level of a villa in the centre. From (2)–(3) we may learn that housing in the centre is more comfortable than housing in the suburbs, which would lead us to conclude from (1) that a villa in the centre would be *luxurious* or *exclusive*. This is a form of extrapolative reasoning, which builds on the premise that analogous changes in the conditions should lead to analogous changes in the conclusions. It can be related to an underlying notion of direction which is defined on the labels of the same class. In the example, for instance, we make the underlying assumption that the change from *basic* to *very-comfortable* goes in the same direction as the change from *luxurious* to *exclusive*.

A more detailed characterization of interpolative and extrapolative inference will be given below. However, it should be clear that in order to automate such inferences, we need a richer form of knowledge than what is available in a classical logical setting, viz. information about betweenness and directionality for labels. In simple domains, we can specify such information by hand. The *Comfort* class, for instance, is essentially the discretization of a linearly ordered numerical domain, hence it suffices to rank the labels. In multi-dimensional domains, however, things are not always so clear. To some extent, a partial description may be manually specified, e.g. we may explicitly assert that the change from a castle to a villa goes in the same direction as the change from a villa to a rowhouse. In large domains, however, it is tedious to provide such specifications, as there is a cubic number of tuples that needs to be considered for betweenness and a quartic number of tuples that needs to be considered for directionality. Moreover, providing this information requires deep knowledge of the considered domain. To cope with this, in this paper, we propose a data-driven approach to acquire the required background knowledge from the web in an automated manner.

In particular, we take advantage of the fact that the notions of betweenness and direction have a clear geometric interpretation, which can be related to Gärdenfors' theory of conceptual spaces [5]. This theory posits that natural properties can be represented as convex regions in a vector space, whose dimen-

sions are called *quality dimensions*, referring to the fact that they correspond to particular qualities (i.e. elementary properties) that objects may exhibit. A typical example are colors, which can be represented in a cognitively meaningful way using three quality dimensions, corresponding to hue, saturation and intensity. By assuming that all labels from the same class can be represented as convex regions in the same conceptual space, the notions of betweenness and direction can be interpreted in their usual, geometric sense. However, there remains the problem of acquiring access to the conceptual spaces representation of the labels. In particular, for most domains, it is not clear what exactly are the quality dimensions, or even how many such dimensions there are. To cope with this, [6] proposes to use multidimensional scaling, which is a well-known family of techniques that can be used to represent a set of objects in a Euclidean space of predefined dimension, starting from similarity judgements for each pair of objects. In particular, the resulting representation is such that, to the best extent possible, two objects are located close to each other in this space iff they were judged to be similar.

In this paper, we continue on this idea, and explore the use of multidimensional scaling to acquire information about the betweenness and directionality of labels, with the aim of supporting interpolative and extrapolative reasoning. To illustrate the proposed techniques, we focus on the music domain. Especially, we explore the possibility of using a purely data-driven approach, starting from tags (i.e. short textual descriptions) that were provided by users of the music recommendation website last.fm[3].

The paper is structured as follows. In the next section, we focus on the idea of building a conceptual spaces representation for a class of labels, using multidimensional scaling. In particular, we illustrate how a conceptual space of music genres can be obtained from publicly available data. Next, Section 3 focuses on interpolative reasoning, showing how betweenness for labels can be derived from a conceptual spaces representation using linear programming. Subsequently, Section 4 discusses extrapolative reasoning, again using a linear programming encoding. Finally, some related work is discussed in Section 5, after which we conclude.

2 Acquiring conceptual representations

As a case study, throughout the paper we focus on the domain of music genres. We may, for instance, consider a knowledge base containing information about what genres are suitable for a particular occasion, e.g. which genres are suitable as background music while working, which are suitable as background music in a bar, which are suitable for dancing, etc. Due to the high number of different genres, it is virtually impossible for such a knowledge base to be complete. For

[3] `http://www.last.fm`, accessed on March 13th, 2011.

instance, Wikipedia mentions hundreds of *popular* music genres[4], and even a few thousand music genres in general[5].

2.1 Obtaining conceptual spaces from tags

For simplicity, we will identify a genre with a set of artists. This means that we are looking for the representation of a conceptual space in which artists correspond to points and genres to regions, and thus that we need similarity judgements for pairs of artists. To obtain these similarity scores, we rely on the music recommendation website last.fm, which among others allows users to assign tags to artists. For each genre in the aforementioned Wikipedia list of popular music genres, we have retrieved the set of artists that were most often tagged with the name of this genre, using the standard last.fm API methods[6]. For a genre g, let A_g be the set of artists that was thus obtained, and let $\mathcal{A} = \bigcup_g A_g$ be the set of all artists that were retrieved. For a tag t, let $count(a,t)$ be the number of times artist a was tagged with tag t. Similarity between artists can then be measured using the following variant of the Jaccard measure:

$$sim(a_1,a_2) = \frac{\sum_t \min(count(a_1,t), count(a_2,t))}{\sum_t \max(count(a_1,t), count(a_2,t))} \tag{4}$$

From the similarity scores, the artists in $\mathcal{A}$ can be mapped to points in a Euclidean space of an arbitrary dimension using multidimensional scaling. For this purpose, we have used the implementation of classical multidimensional scaling of the MDSJ java library[7].

Figure 1 depicts, for a selected number of genres, the locations of the artists that were obtained after multidimensional scaling to two dimensions. For clarity, we display the names of the corresponding genres, rather than the names of the artists themselves. Note that two dimensions is clearly not enough to capture all relevant aspects of music genres. We only use two dimensions to visualize some aspects of the data set, and we will use larger numbers of dimensions below. In general, the larger the number of dimensions, the better the Euclidean space representation will be in accordance with the similarity judgements. On the other hand, by choosing the number of dimensions too high, relevant structure may be lost.

2.2 Typicality

When looking at the genres in Figure 1, we notice that genres tend to consist of a rather compact core, where most artists of the genre are located, together

[4] `http://en.wikipedia.org/wiki/List_of_popular_music_genres`, accessed on March 10th, 2011.

[5] `http://en.wikipedia.org/wiki/List_of_music_styles`, accessed on March 10th, 2011.

[6] `http://www.last.fm/api`

[7] `http://www.inf.uni-konstanz.de/algo/software/mdsj/`, accessed on March 11th, 2011.

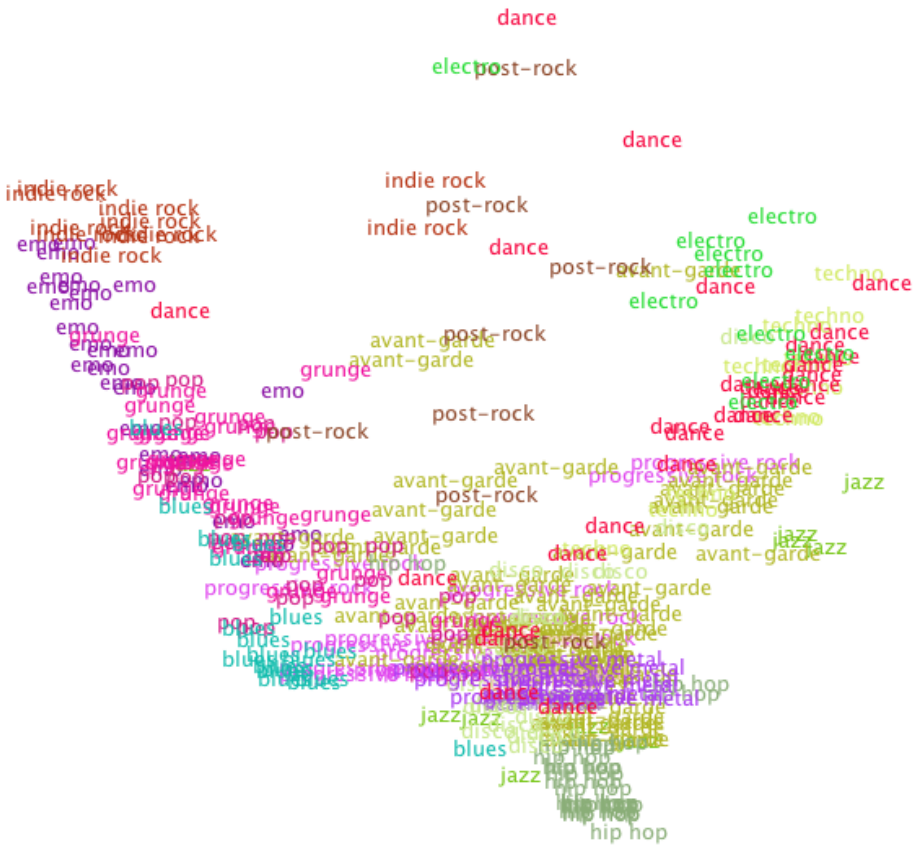

Fig. 1. Locations of artists after multidimensional scaling with two dimensions.

with a number of outliers that are located further away. While to some extent this is due to noisy input (e.g. the fact that measuring similarity in terms of tags is not a perfect method), it is also the case that each genre indeed has artists that are typical of the genre, as well as artists that are more borderline. It is to be expected that the typical artists are those that belong to the core of the geometric representation. Note that this idea that typicality can be identified with geometric centrality is a common assumption in the theory of conceptual spaces [5].

The notion of typicality plays an important role in commonsense reasoning, as it is often the case that rules only apply to typical situations. Consider for instance a rule such as

$$jazz(x) \rightarrow dissonant(x)$$

which asserts that one of the characteristics of jazz music is its use of dissonant chords. While this is true for most jazz music, it is not hard to imagine that there might be some exceptional jazz artists who adhere to a consonant style.

It seems natural to represent a given genre as the convex hull of the location of its artists. To represent the core of the genre, i.e. the set of its typical instances, we will consider the convex hull of the locations of the most central artists. Specifically, for each genre we calculate which is the most central artist, i.e. we determine the medoid c_g of a genre g as follows:

$$c_g = \arg\min_{a \in A_g} \sum_{a' \in A_g} dist(a, a')$$

where $dist(a, a')$ is the Euclidean distance between the locations of artists a and a' that were obtained after multidimensional scaling. Then we may geometrically represent a genre g as the convex hull[8] of the locations of those artists that are closest to c_g. In the remainder of this paper, we will consider the sets A_g^{25}, A_g^{50}, A_g^{75} and A_g^{100} containing the 25%, 50%, 75% and 100% closest artists to c_g.

Figure 2 shows the bounding boxes of the sets A_g^{75}, for each of the genres from Figure 1. What is particularly noticeable is that there are some genres that have a compact representation (e.g. hip hop, death metal, indie rock) and others that are quite dispersed (e.g. dance, post-rock, jazz). The compactness of these representations appears to be related to the variety in styles that the genre may encompass: post-rock, for instance, encompasses aspects of ambient, jazz, and electronica, but using rock music instruments[9]. Figure 2 also illustrates the limitations of a two-dimensional representation, as e.g. disco and black metal are incorrectly represented as subsets of jazz.

3 Interpolative inference

In [12] we propose a form of interpolative inference which is centered around the notion of betweenness. Intuitively, we say that a label b is between the labels a and c if every relevant feature which is shared among the labels a and c is also present in b[10]. If we then know that the rules $a(x) \rightarrow u(x)$ and

[8] Instead of using the convex hull, [6] proposes to use a generalized form of Voronoi diagrams. This, however, requires that the labels of one class are jointly exhaustive and pairwise disjoint, which seems too strong an assumption in the case of music genres (e.g. there may be sub-genres and crossover-genres). While using the convex hull leads to a conservative representation, underestimating the extent of a region, this does not seem to pose any problems in the considered setting, where geometrical representations are largely restricted to the most typical artists anyway.

[9] `http://en.wikipedia.org/wiki/Post-rock`, accessed March 11th, 2011.

[10] This intuition is fairly in agreement with a formal view of analogical proportion that has recently been proposed [9]. Indeed, stating a logical proportion of the form a is to b as c is to d amounts to express that a differs from b as c differs from d and that b differs from a as d differs from c. This corresponds to the propositional logic expression $((a \wedge \neg b) \equiv (c \wedge \neg d)) \wedge ((b \wedge \neg a) \equiv (d \wedge \neg c))$. This expression is logically equivalent to $((a \wedge d) \equiv (b \wedge c)) \wedge ((\neg a \wedge \neg d) \equiv (\neg b \wedge \neg c))$ [9]. In the particular case of the continuous logical proportion a is to b as b is to c, which corresponds to the idea of having b *between* a and c, the latter logical expression reduces to

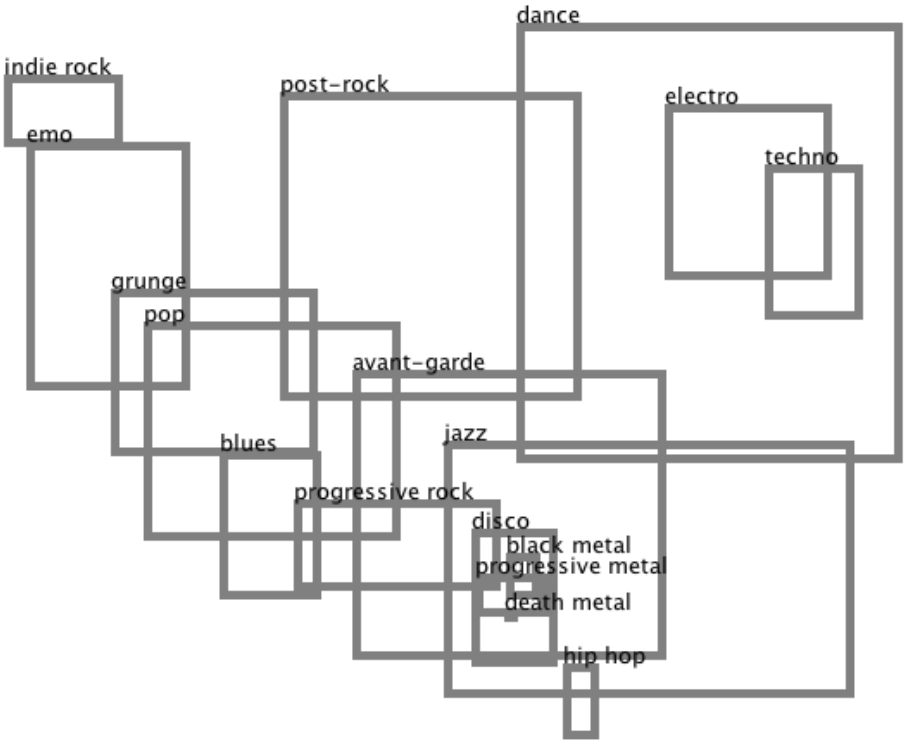

Fig. 2. Bounding boxes of the sets A_g^{75}, containing the 75% most central artists of genre g.

$c(x) \rightarrow w(x)$ are both valid rules, interpolative reasoning allows us to conclude that when $b(x)$ holds for an object x, there exists some label v which is between the labels u and w such that $v(x)$ holds. This idea can readily be extended to betweenness for more than two labels. It can also be extended to cope with rules with conjunctions or disjunctions in the antecedent and consequent, but the details are more technical and fall outside the scope of this paper.

Example 2. Let us consider the following information about music genres:

$$\textit{samba}(x) \rightarrow \textit{standard-harmony}(x)$$
$$\textit{jazz}(x) \rightarrow \textit{complex-harmony}(x)$$

Knowing that *bossa nova* is between *samba* and *jazz*, we conclude that the complexity of harmonizations in *bossa nova* music is between standard and complex.

$((a \wedge c) \equiv b) \wedge ((\neg a \wedge \neg c) \equiv \neg b)$. This amounts to say, both positively and negatively, that what a and c have in common, b has it also (and conversely, in the case of analogical proportions).

From a practical point of view, one of the main problems is to decide for a set of labels $a_1, ..., a_n$ of the same class, which other labels can be considered to be between them. Using the representations that were obtained in Section 2, this becomes a matter of geometric compution, i.e. genre g is between genres $g_1, ..., g_n$ if the artists in A_g are located in the convex hull of the locations of the artists in $A_{g_1} \cup ... \cup A_{g_n}$. Ideally, we want interpolation also to be meaningful when it is applied to default rules, i.e. rules which only hold for typical instances. Thus we are interested in discovering instances of betweenness which remain meaningful when genres are restricted to their most typical artists. More precisely, let us consider $bet^\lambda(g, \{g_1, ..., g_n\})$ as a measure of the extent to which g is between $g_1, ..., g_n$ at a typicality level of $\lambda\%$, i.e.

$$bet^\lambda(g, \{g_1, ..., g_n\}) = \frac{|\{a \in A_g^\lambda : a \in cvx(A_{g_1}^\lambda \cup ... \cup A_{g_n}^\lambda)\}|}{|A_g^\lambda|}$$

where *cvx* denotes the convex hull, and we identify artists with the corresponding points in the Euclidean space that was obtained by multidimensional scaling. The score $bet(g, \{g_1, ..., g_n\})$ then provides an overall estimate of the appropriateness to consider g as being between $g_1, ..., g_n$:

$$bet(g, \{g_1, ..., g_n\}) = \frac{1}{4} \cdot \sum_{\lambda \in \{25,50,75,100\}} bet^\lambda(g, \{g_1, ..., g_n\})$$

Unfortunately, deriving an explicit representation of the convex hull of a set of points in higher dimensions is a computationally expensive task, and may require an exponential amount of time and space. However, we do not actually need this representation; we only need a procedure which allows us to decide whether a point p is in the convex hull of a set of points $\{q_1, ..., q_m\}$. This can be checked in polynomial time using linear programming solvers. In particular, let us write p^i and q_j^i to denote the i^{th} coordinate of p and q_j respectively, and assume that there are d dimensions in total. Then we consider the following set of linear (in)equalities in the variables $\lambda_1, ..., \lambda_m$:

$$\begin{aligned} \Gamma = &\{\lambda_1 q_1^i + ... + \lambda_m q_m^i = p^i \ : \ 1 \leq i \leq d\} \\ &\cup \{\lambda_j \geq 0 \ : \ 1 \leq j \leq m\} \cup \{\lambda_1 + ... + \lambda_m = 1\} \end{aligned}$$

Then clearly p is in the convex hull of $\{q_1, ..., q_m\}$ iff Γ has a solution.

Table 1 shows the genres that we found to be between *jazz* and *samba* music, considering multidimensional scaling in 2 to 6 dimensions. As expected, in 2 dimensions, the results are not always reliable, e.g., it is hard to justify that a genre such as *deathrock* should conceptually be between *jazz* and *samba*. However, as soon as 3 or 4 dimensions are used, most of the results seem to be reasonable. Apart from some reranking, no major differences are seen among the results for 4, 5 and 6 dimensions, although the absolute scores drop substantially.

Table 2 provides an example where more than two genres are initially given. We may consider, for instance, a user who needs to provide a music recommendation system with the genres she likes. From an initial seed of genres, the

2 dimensions	3 dimensions	4 dimensions	5 dimensions	6 dimensions
big band : 0.92	jazz fusion : 0.77	big band : 0.89	big band : 0.54	big band : 0.44
vocal jazz : 0.85	big band : 0.68	swing : 0.64	smooth jazz : 0.48	swing : 0.33
smooth jazz : 0.82	smooth jazz : 0.67	smooth jazz : 0.62	swing : 0.44	smooth jazz : 0.20
bossa nova : 0.79	free jazz : 0.67	jazz fusion : 0.45	bossa nova : 0.29	free jazz : 0.13
jazz fusion : 0.79	swing : 0.57	bossa nova : 0.44	jazz fusion : 0.27	vocal jazz : 0.12
swing : 0.78	salsa : 0.43	free jazz : 0.42	vocal jazz : 0.15	bossa nova : 0.10
salsa : 0.63	bossa nova : 0.41	vocal jazz : 0.29	nu jazz : 0.15	jazz fusion : 0.09
dancehall : 0.63	vocal jazz : 0.39	vocal : 0.19	free jazz : 0.10	easy listening : 0.08
free jazz : 0.61	disco : 0.25	easy listening : 0.18	acid jazz : 0.10	vocal : 0.06
deathrock : 0.53	bluegrass : 0.23	salsa : 0.16	salsa : 0.09	salsa : 0.05

Table 1. Music genres g that are between *jazz* and *samba*, with the corresponding values of $bet(g, \{jazz, samba\})$.

2 dimensions	3 dimensions	4 dimensions	5 dimensions	6 dimensions
grunge : 1.0	emo : 1.0	pop rock : 0.68	pop rock : 0.54	emo : 0.50
pop punk : 1.0	power pop : 0.98	emo : 0.67	emo : 0.41	pop punk : 0.43
emo : 1.0	pop rock : 0.98	britpop : 0.58	britpop : 0.39	britpop : 0.34
post-grunge : 0.99	grunge : 0.98	grunge : 0.56	post-grunge : 0.31	pop rock : 0.34
pop rock : 0.97	pop punk : 0.88	pop punk : 0.54	grunge : 0.28	grunge : 0.32
math rock : 0.97	punk rock : 0.85	power pop : 0.50	pop punk : 0.25	power pop : 0.29
britpop : 0.95	folk rock : 0.74	post-grunge : 0.39	power pop : 0.25	soft rock : 0.14
post-punk : 0.95	new wave : 0.72	new wave : 0.36	uplifting : 0.19	glam rock : 0.12
indie folk : 0.95	britpop : 0.72	soft rock : 0.34	new wave : 0.17	punk rock : 0.09
new wave : 0.95	garage rock : 0.69	garage rock : 0.31	soft rock : 0.15	uplifting : 0.07

Table 2. Music genres g that are between *indie rock*, *pop*, *alternative rock* and *rock*, with the corresponding values of $bet(g, \{indie\ rock, pop, alternative\ rock, rock\})$.

system may then try to build a more complete user profile in an automated fashion. In the case of Table 2, the user has provided four genres, viz. *indie rock*, *pop*, *alternative rock*, and *rock*. Intuitively, we would expect to find a variety of sub-genres of rock music, including both mainstream and niche genres. In this case, the results appear already reasonable when using 2 dimensions, with the exception perhaps of *indie folk*. In general, compared to Table 1, the scores that are obtained in Table 2 are higher. In part, this is due to the fact that there is a larger number of genres that is clearly relevant. However, the fact that four genres were provided instead of two, also seems to make the results more robust to idiosyncrasies of the multidimensional scaling algorithm.

It is hard to provide a quantitative evaluation of the performance of our method, as the question as to which results should be considered correct is highly subjective and dependent on the application context which is envisioned. However, results such as the ones in Table 1 and 2 do suggest that the data-driven approach to commonsense reasoning which we put forward in the paper is indeed feasible. Evaluating this approach in an end-to-end system — say, a music recommendation engine — will be a topic of future work.

4 Extrapolative inference

The extrapolative inferences that we consider are centered around the notion of *direction of change*. Considering four labels a, b, c and d, we are interested in knowing whether the transition from a to b affects the same features as the transition from c to d. For instance, we may consider that the transition from *hard rock* to *progressive rock* is in the same spirit as the transition from *heavy metal* to *progressive metal*. The idea is that from the rules $a(x) \rightarrow u(x)$, $b(x) \rightarrow v(x)$ and $c(x) \rightarrow w(x)$ we want to conclude something about $d(x)$. Knowing that the transition from a to b goes in the same direction as the transition from c to d, we may conclude that when $d(x)$ is the case, then some $z(x)$ holds, such that the change from u to v is in the same direction as the change from w to z.

Example 3. Consider the following rule base:

$$\begin{aligned} \textit{hardrock}(x) &\rightarrow \textit{mainstream}(x) \\ \textit{progrock}(x) &\rightarrow \textit{borderline-mainstream}(x) \\ \textit{heavymetal}(x) &\rightarrow \textit{borderline-mainstream}(x) \end{aligned}$$

which encodes that hard rock artists can be considered to belong to a mainstream genre, while progressive rock artists and heavy metal artists are on the boundary between mainstream and niche music. Using extrapolative reasoning, we conclude that progressive metal artists are either borderline-mainstream or are in a niche, i.e.

$$\textit{progmetal}(x) \rightarrow \textit{borderline-mainstream}(x) \vee \textit{niche}(x)$$

The notion of direction of change can be defined more precisely in terms of conceptual spaces. If A, B, C and D are the convex sets that represent the

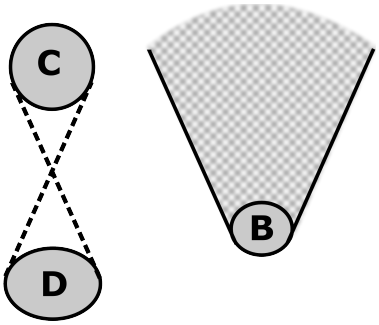

Fig. 3. Geometric representation of constraint (5).

labels a, b, c and d in a conceptual space, we say that the change from a to b goes in the same direction as the change from c to d if each of the following conditions is satisfied:

$$\forall p \in A\,.\,\exists q \in B, r \in C, s \in D\,.\,\exists \lambda > 0\,.\,\overrightarrow{pq} = \lambda \cdot \overrightarrow{rs} \tag{5}$$

$$\forall q \in B\,.\,\exists p \in A, r \in C, s \in D\,.\,\exists \lambda > 0\,.\,\overrightarrow{pq} = \lambda \cdot \overrightarrow{rs} \tag{6}$$

$$\forall r \in C\,.\,\exists p \in A, q \in B, s \in D\,.\,\exists \lambda > 0\,.\,\overrightarrow{pq} = \lambda \cdot \overrightarrow{rs} \tag{7}$$

$$\forall s \in D\,.\,\exists p \in A, q \in B, r \in C\,.\,\exists \lambda > 0\,.\,\overrightarrow{pq} = \lambda \cdot \overrightarrow{rs} \tag{8}$$

The intuition behind (5) is illustrated in Figure 3. Specifically, it holds that (5) is satisfied iff A is included in the shaded region above region B. In fact, each of these four conditions supports some forms of extrapolative inference. In particular, (5) is needed to extrapolate information concerning labels b, c and d to label a, while (6) is needed to extrapolate information concerning a, c and d to label b, etc. Also note that this notion of direction of change is strongly related to reasoning by analogy [10]. In fact, we say that a, b, c and d are in an analogical proportion, i.e. that "a is to b what c is to d", if the *direction* of the change from a to b is the same as the direction of the change from c to d, and moreover, the *amount* of change from a to b is the same as the amount of change from c to d. This means that the idea of analogical proportions could be formalized by insisting that $\lambda = 1$ in (5)–(8), which is well in agreement with the standard parallelogram view of analogical proportions.

As for betweenness, a linear programming formulation can be used to check whether (5)–(8) are satisfied. For example, it holds that (5) is satisfied for a point $p \in A$ iff the following system of linear (in)equalities has a solution:

$$\begin{aligned}\Gamma = &\{\lambda_1^b q_1^i + ... + \lambda_m^b q_m^i + \lambda_1^d s_1^i + ... + \lambda_l^d s_l^i - (\lambda_1^c r_1^i + ... + \lambda_k^c r_k^i) = p^i : 1 \leq i \leq d\} \\ &\cup \{\lambda_j^b \geq 0 : 1 \leq j \leq m\} \cup \{\lambda_j^c \geq 0 : 1 \leq j \leq k\} \cup \{\lambda_j^d \geq 0 : 1 \leq j \leq l\} \\ &\cup \{\lambda_1^b + ... + \lambda_m^b = 1\} \cup \{\lambda_1^c + ... + \lambda_k^c - (\lambda_1^d + ... + \lambda_l^d) = 0\}\end{aligned}$$

where $B = cvx(\{q_1, ..., q_m\})$, $C = cvx(\{r_1, ..., r_k\})$ and $D = cvx(\{s_1, ..., s_l\})$, and e.g. q_j^i is the i^{th} coordinate of point q_j as before. To see the relationship between

the (in)equalities in Γ and (5), it is useful to note that the variables λ_j^b, λ_j^b and λ_j^a are used to find the points q, r and s from (5). In particular:

$$q = \lambda_1^b q_1 + ... + \lambda_m^b q_m$$
$$r = \frac{\lambda_1^c}{\lambda} r_1 + ... + \frac{\lambda_k^c}{\lambda} r_k$$
$$s = \frac{\lambda_1^d}{\lambda} s_1 + ... + \frac{\lambda_l^d}{\lambda} s_l$$

Furthermore note that we need to insist that $\sum_i \lambda_i^b = 1$ and $\sum_i \lambda_i^c = \sum_i \lambda_i^d = \lambda$ to ensure that q, r and s are in the respective convex hulls.

Based on this linear programming implementation, we can calculate the following score, measuring to what extent the transition from genre g_1 to g_2 is in the same direction as the transition from genre g_3 to genre g_4:

$$dir(g_1, g_2; g_3, g_4) = \frac{1}{4} \cdot \sum_{\lambda \in \{25,50,75,100\}} \Big(\min_{x \in \{1,2,3,4\}} \frac{|\{p \,:\, p \in A_{g_1}^\lambda, cond_x^\lambda(p)\}|}{|A_{g_1}^\lambda|} \Big)$$

where $cond_1^\lambda(p_j)$ checks whether (5) is satisfied for point p_j, and $cond_2^\lambda$, $cond_3^\lambda$ and $cond_4^\lambda$ correspond in the same way to (6)–(8).

Tables 3 and 4 show those pairs of genres (g_1, g_2) for which the transition from g_1 to g_2 goes in the same direction as the transition from *hard rock* to *progressive rock* (Table 3) and the transition from *indie rock* to *pop* (Table 4). As the score $dir(g_1, g_2; g_3, g_4)$ is trivially high when there is a large overlap between g_1 and g_2, or between g_3 and g_4, pairs of overlapping genres were excluded. Note that the (geometric) overlap of genres g_1 and g_2 can be evaluated using $bet(g_1, \{g_2\})$ and $bet(g_2, \{g_1\})$.

Compared to the results we obtained for betweenness in Tables 1 and 2, the scores in Tables 3 and 4 are considerably higher. While meaningful results are found overall, in 4 and 6 dimensions some intuitively incorrect pairs are found such as (*glam rock*, *vocal*) and (*rapcore*, *uplifting*). For 8 dimensions, however, only relevant pairs are found. Similarly, in Table 4, mostly relevant results are found, although it is not clear whether a pair such as (*emo*, *new wave*), which is found when using 4 dimensions, should be considered correct.

Overall, we may conclude that useful results are obtained, provided that a sufficiently high number of dimensions is chosen. There are at least two reasons why we seem to need a larger number of dimensions here, than for betweenness. First, as we consider pairs of genres here, there are considerably more candidates (quadratic in the number of genres instead of linear), hence there is a larger risk that the regularities found are due to chance. By increasing the number of dimensions, any remaining structure is more likely to be intrinsic. Second, whenever the distance between the geometric representation of two genres g_1 and g_2 is small, relative to their sizes, the constraints that two genres g_3 and g_4 should define the same directions is easier to satisfy than the betweenness constraint from Section 3. As a result, the influence of outliers and noise is also potentially higher when looking for parallel distances.

4 dimensions	6 dimensions	8 dimensions
(hard rock,southern rock) : 0.78	(rapcore,uplifting) : 0.61	(glam rock,progressive rock) : 0.45
(nu metal,post-hardcore) : 0.77	(southern rock,psychedelic rock) : 0.61	(nu metal,progressive rock) : 0.44
(pop rock,dream pop) : 0.76	(southern rock,space rock) : 0.59	(groove metal,mathcore) : 0.42
(hard rock,psychedelic rock) : 0.75	(southern rock,progressive rock) : 0.59	(southern rock,progressive rock) : 0.37
(groove metal,doom metal) : 0.74	(power pop,dream pop) : 0.58	(groove metal,progressive metal) : 0.35
(pop punk,dream pop) : 0.74	(rapcore,dream pop) : 0.57	(glam rock,space rock) : 0.35
(emo,britpop) : 0.74	(glam rock,progressive rock) : 0.57	(screamo,space rock) : 0.34
(glam rock,post-punk) : 0.73	(punk rock,garage rock) : 0.57	(post-hardcore,space rock) : 0.34
(glam rock,vocal) : 0.73	(groove metal,mathcore) : 0.57	(nu metal,math rock) : 0.33
(pop rock,lo-fi) : 0.72	(rapcore,lo-fi) : 0.57	(groove metal,doom metal) : 0.33

Table 3. Pairs of music genres (g_1, g_2) such that the transition from g_1 to g_2 goes in the same direction as the transition from *hard rock* to *progressive rock*.

4 dimensions	6 dimensions	8 dimensions
(britpop,pop) : 0.75	(britpop,pop) : 0.61	(britpop,pop) : 0.56
(punk rock,glam rock) : 0.56	(indie pop,pop) : 0.47	(indie pop,pop) : 0.43
(lo-fi,disco) : 0.54	(indie rock,pop rock) : 0.42	(emo,pop rock) : 0.30
(lo-fi,easy listening) : 0.54	(indie rock,britpop) : 0.37	(indie rock,pop rock) : 0.27
(emo,soft rock) : 0.53	(emo,glam rock) : 0.36	(emo,pop punk) : 0.25
(emo,new wave) : 0.53	(emo,pop rock) : 0.36	(britpop,pop rock) : 0.24
(emo,power pop) : 0.52	(power pop,pop) : 0.36	(indie rock,britpop) : 0.22
(britpop,disco) : 0.51	(dream pop,disco) : 0.35	(pop punk,pop rock) : 0.21
(indie rock,pop rock) : 0.51	(emo,soft rock) : 0.34	(uplifting,pop) : 0.20
(emo,pop punk) : 0.50	(lo-fi,disco) : 0.33	(garage rock,rock and roll) : 0.19

Table 4. Pairs of music genres (g_1, g_2) such that the transition from g_1 to g_2 goes in the same direction as the transition from *indie rock* to *pop*.

5 Related work

The idea of interpolation and extrapolation of knowledge has already been studied in a number of different settings. Interpolation has extensively been studied in the context of fuzzy set theory [11, 2, 1], although predominantly in numerical settings. The main idea underlying such methods is that a rule such as "if a then b" is interpreted as "the more we are in a situation similar to a, the more it holds that we are in a situation similar to b". Thus, fuzzy set based methods also start from numerical similarity information, as we did in Section 2, but they use such information in a more direct way. Extrapolation of knowledge is also studied in [4], in the restricted setting of time-stamped propositional knowledge. In particular, an approach is introduced to extrapolate information about the beliefs that are held at a given moment in time to beliefs about other time points.

More generally, the kind of interpolative and extrapolative reasoning patterns that we have considered are motivated by the idea that in absence of any other information, it is reasonable to assume that when completing available knowledge, we should not introduce any irregularities. Starting from a similar motivation, [7] studies the problem of ranking a set of alternatives according to a given set of constraints. In [10], analogical proportions are used as the basis for extrapolating from known cases in a machine learning perspective.

As taxonomies can be identified with sets of rules, the work in this paper is also somewhat related to data-driven approaches for refining taxonomies. For example, [3] uses formal concept analysis to introduce intermediary labels when merging different taxonomies.

Apart from the work on conceptual spaces, the idea of assuming a spatial representation to reason about concepts also underlies [8], where an approach to integrate heterogeneous databases is proposed based on spatial relations between concepts. The use of semantic background information about the relation between different labels also underlies a recent proposal for merging inconsistent propositional knowledge bases [13].

6 Conclusions

We have presented the outline of a purely data-driven approach to interpolative and extrapolative reasoning. Starting from pairwise similarity measurements for the instances of the domain of discourse, a geometric representation of these instances is obtained as points in a conceptual space using multidimensional scaling. Properties or concepts, denoted by natural language labels and identified with sets of instances, can then be represented as convex regions in this conceptual space. Rather than constructing these convex regions explicitly, which may require an exponential amount of space, we rely on a linear programming formulation to derive information about the spatial relations that hold between the (unknown) geometric representations of labels.

Although this general idea is applicable to any domain where similarity can be measured, we have focused specifically on the domain of music genres to

illustrate our proposed method. The examples that we have provided illustrate that good results may be obtained, provided that (i) care is taken to alleviate the effect of outliers/noise, and (ii) a sufficiently high number of dimensions is chosen.

Acknowledgments Steven Schockaert was funded as a postdoctoral fellow by the Research Foundation – Flanders (FWO).

References

1. B. Bouchon-Meunier, F. Esteva, L. Godo, M. Rifqi, and S. Sandri. A principled approach to fuzzy rule base interpolation using similarity relations. In *Proc. of the EUSFLAT-LFA Joint Conference, Barcelona*, pages 757–763.
2. D. Dubois, H. Prade, F. Esteva, P. Garcia, and L. Godo. A logical approach to interpolation based on similarity relations. *International Journal of Approximate Reasoning*, 17(1):1 – 36, 1997.
3. F. Dupin de Saint-Cyr, R. Jeansoulin, and H. Prade. Spatial information fusion: Coping with uncertainty in conceptual structures. In *ICCS Supplement*, pages 66–74, 2008.
4. F. Dupin de Saint-Cyr and J. Lang. Belief extrapolation (or how to reason about observations and unpredicted change). *Artificial Intelligence*, 175(2):760 – 790, 2011.
5. P. Gärdenfors. *Conceptual Spaces: The Geometry of Thought*. MIT Press, 2000.
6. P. Gardenfors and M. Williams. Reasoning about categories in conceptual spaces. In *International Joint Conference on Artificial Intelligence*, pages 385–392, 2001.
7. R. Gérard, S. Kaci, and H. Prade. Ranking alternatives on the basis of generic constraints and examples: a possibilistic approach. In *Int. Joint Conf. on Artifical intelligence*, pages 393–398, 2007.
8. F. Lehmann and A. G. Cohn. The EGG/YOLK reliability hierarchy: semantic data integration using sorts with prototypes. In *Int. Conf. on Information and Knowledge Management*, pages 272–279, 1994.
9. L. Miclet and H. Prade. Handling analogical proportions in classical logic and fuzzy logics settings. In C. Sossai and G. Chemello, editors, *Proc. 10th Europ. Conf. on Symbolic and Quantitative Approaches to Reasoning with Uncertainty (ECSQARU'09), Verona, Jul. 1-3*, volume 5590 of *LNCS*, pages 638–650. Springer, 2009.
10. H. Prade and G. Richard. Reasoning with logical proportions. In *International Conference on Principles of Knowledge Representation and Reasoning (KR)*, pages 545–555, 2010.
11. E. Ruspini. On the semantics of fuzzy logic. *International Journal of Approximate Reasoning*, 5:45–88, 1991.
12. S. Schockaert and H. Prade. Qualitative reasoning about incomplete categorization rules based on interpolation and extrapolation in conceptual spaces. In *Proceedings of the Fifth International Conference on Scalable Uncertainty Management*, 2011.
13. S. Schockaert and H. Prade. Solving conflicts in information merging by a flexible interpretation of atomic propositions. *Artificial Intelligence*, 175:1815–1855, 2011.