Automatic extraction of ingredient’s substitutes

Corrado Boscarino  
VU Amsterdam  
De Boelelaan 1081A  
1081 HV Amsterdam NL  
c.boscarino@vu.nl

Nicole J.J.P. Koenderink  
Wageningen UR FBR  
Bornse Weilanden 9  
6708 WG Wageningen NL  
Nicole.Koenderink@wur.nl

Vladimir Nedović  
Flavourspace  
Mauritskade 55C  
1092 AD Amsterdam NL  
v@flavourspace.com

Jan L. Top  
Wageningen UR FBR  
VU Amsterdam  
De Boelelaan 1081A  
1081 HV Amsterdam NL  
jan.top@wur.nl

Abstract
Expert advice on how ingredients can be replaced in recipes is widely available on-line. However, these are general substitution rules, which do not take into account contextual factors such as culture, sensory perception, season, etc. We aim at tuning general rules to particular recipes. From an on-line food encyclopedia we extract explicit substitution rules. We also consider implicit substitution rules, derived by the categorisations in the same source. By applying Latent Dirichlet Allocation (LDA) onto a crawled dataset, we rank ingredients based on their likelihood of being interchangeable, given a recipe. The results show that our statistical approach can approximate manual judgments.

Author Keywords
conference publications LDA, recipe recommendation, ingredients substitution

Introduction
The role of food in environmental sustainability is a highly contested notion. Bringing about a radical change in food production and consumption patterns is a top priority of both governmental [8] and private organisations [10]. Promoting these community transitions through user empowerment and lifestyle change is an especially lively area of research in HCI [3].
A growing amount of everyday food practices are taking place on-line. Food blogs, collections of recipes submitted by end-users and commercial repositories are widely used for planning a meal. Awareness about healthy, seasonal, locally and on the long term sustainable produced food leads users of on-line services to seek for categorisations that capture their preferred features or for the means of adapting existing recipes to their needs. HCI technologies have been traditionally oriented towards compensating some deficiency at the user's side [9]. Lack of nutrition knowledge, skills or information for finding suitable recipes motivated the design of systems for visualising the nutritional content of ingredients, for browsing and selecting recipes, and for categorising recipes based on user generated facets [16, 17]. Personalisation is achieved by developing models of a dataset, for example by applying collaborative filtering algorithms [7] or ingredient networks [15].

Extensive qualitative studies show that domestic cooking skills are, however, surprisingly complex [14]. Failing to take into account contextual features, such as occasion, food availability, taste and preferences might prevent successful and meaningful interactions [4, 5] and ultimately behavioural change.

While well designed food blogs 1 and recipe repositories 2 provide enough contextual information for assessing the suitability of a meal to user requirements, expert advice can still fail to reach its intended audience [1]. For this work we consider the Cook's Thesaurus 3 (CT), a cooking encyclopedia that covers thousands of ingredients. For many ingredients the CT suggests possible substitution rules; information on how the rule should be applied and what consequences a rule has on the resulting preparation is often provided between brackets. For example cow's milk substitutes are goat's milk, almond milk, oat milk, soy milk, water and sour cream:

Substitutes: goat’s milk (This is easier to digest that cow’s milk. Fresh goat’s milk is a good all-purpose substitute for cow’s milk, but non-refrigerated forms have an unpleasant tangy, barnyard flavor that overpowers subtly-flavored dishes.) OR oat milk (This is more expensive and less nutritious, but a good all-purpose substitute for milk in cooking. A good choice for vegetarians who object to the animal exploitation involved in the production of cow’s or goat’s milk.) OR buttermilk (This is especially good in pancakes or waffles. If using in a baked good, 1 C milk = 1 C buttermilk - 2 teaspoons baking powder + 1/2 teaspoon baking soda) OR soy milk (This is more expensive and (unless fortified) less nutritious than ordinary milk. It has a nutty flavor and turns beige when cooked. It works well in most baked goods, but it’s a risky substitute in savory dishes.) OR rice milk (This is more expensive and, unless fortified, less nutritious. It’s great for making desserts, but it’s too sweet for savory dishes) OR almond milk (This also is more expensive and, unless fortified, less nutritious. It’s very sweet; use it in desserts only) OR water (makes eggs creamier in scrambled eggs, gives bread a coarser texture and a lighter crust) OR (in baked goods) sour cream (This makes baked goods more tender and moist. For each cup of milk you replace, subtract one teaspoon of baking powder and up to three tablespoons of fat from the recipe and add one cup of sour cream plus 1/2 teaspoon of baking soda.)

These substitution rules are based on culinary knowledge as biochemical categorisations might not always be effective: whereas the first two items are proper types of milk, others are not even dairy products and felicitous substitution of milk with water is strongly application specific. We assume that an expert developed these rules based on her culinary knowledge, a set of recipes hidden to the CT users, who must then evaluate for each recipe whether or not the rule’s conditions apply. Moreover, the application range of the rules might exceed that initial set of recipes: there are possibly many recipes upon which the substitution rules can be applied, but were not anticipated by experts. We can also derive implicit substitution rules from the CT tree structure. We assume that ingredients that are included in the same category, e.g. types of cheeses, root vegetables, etc. are potentially interchangeable. Since food ontologies can improve recipe retrieval [6], we also derived an ingredient ontology based on CT. Additional interchangeable ingredients are implicitly defined by having at least a common parent.

In this work we investigate whether the cognitive gap

---

1http://www.saveur.com/content/best-food-blog-awards-vote
2http://www.food.com/recipes#
3http://www.foodsubs.com/
between authors and end-users, that is the gap between what an author meant by writing a document, including its possible relevant contexts, and what information seekers mean [11] can be re-built at the system’s side using additional background knowledge. In case of the substitution rules in the CT we aim at ranking ingredients based on their effectiveness as substitutes, given a recipe. While the CT is geared towards American users, we gather recipes from a well known Dutch grocery market company, which also maintains a recipe repository. We applied a statistical analysis on the recipe dataset, deriving an alternative representation of the document in a topic space by means of Latent Dirichlet Allocation (LDA) [2]. We investigated whether the topic structure can bridge the cognitive gap, identifying the most probable substitutes for a given ingredient in a recipe.

We qualitatively evaluated the potential of this method. From CT, we selected few substitution rules that are commonly perceived as difficult and highly context dependent and we ranked the ingredients according to their likelihood of being exchangeable for a particular recipe. Our results show that our approach can compete with human judgments.

**Experimental setting**

We derived a set of rules both implicitly from the CT tree structure and explicitly from the *Substitutes* clauses, each separated by the keyword OR; we recorded the text between brackets for evaluation purposes. We investigated substitute assignment in context. Intuitively, we aim at ranking an ingredient such as Mascarpone higher than, say Boursin, which are both instances of the Fresh Cheese subclass of Cheese, when we detect that a recipe belongs to a ‘pie-baking’ context. Or, for example, we would like to rank Sesam Oil higher than Olive Oil, which are both instances of the Baking Fats class, when we detect an ‘asian cuisine’ context; in the latter context we also aim at ranking shiitake higher in the explicit rule:

```
portobello mushroom
Substitutes: cremini (smaller) OR matsutake (for grilling) OR porcini (for grilling)
OR shiitake OR white mushroom OR oyster mushrooms
```

**The Albert Heijn (AH) dataset**

We seek suitable contexts for the substitution rules in a dataset obtained by crawling 12515 recipes from a Dutch grocery market website using Httrack with default parameters. For this work we reduced the recipes, which contain also cooking procedures, amount of ingredients, among other information, to simple lists of ingredients. Each recipe becomes a point in the 7172-dimensional space of unique ingredients that appear in the AH dataset.

Manual inspection revealed that the crawled recipe ingredients were far from uniformly specified; the original vocabulary can be greatly simplified. Different descriptions were used for the same ingredient, e.g. appeltjes and kleine appels (small apples), kookaardappelen and kookaardappels (potatoes). Some ingredients occurred both in plural and singular form, e.g. abrikoos and abrikozen (apricot). Since our substitution rules were mainly focussed on preserving taste, ingredients that were embellished with adjectives without meaning for the taste of the ingredient, such as free range, organic, pre-cut, thick, or a country of origin can be considered superfluous for our task. Examples include scharrelei-advocaat (eggnog of free range eggs), biologische ham (organic ham), gesneden bladspinazie (pre-cut spinach) or Hollandse aardbeien (Dutch strawberries).

We manually map the original ingredient space to a more

---

4http://www.ah.nl/allerhande/recept, example recipes in this paper are translated by the authors.
5http://www.httrack.com/
uniform list. We maintain per product either the plural or the singular form, we remove all adjectives that are irrelevant for taste, and we choose one spelling per product. We also removed from the vocabulary all ingredients that appears in only one recipe, yielding a reduced space of 2199 unique ingredients.

**Context representation**

For most applications, using the full data representation space is unfeasible. We can reduce its dimensionality by exploiting the strong dependency between words in meaningful documents, such as in food recipes. We apply Latent Dirichlet Allocation (LDA), which assumes that a restricted set of \( k \) topics can make sense of the data. If recipes are reduced to lists of ingredients, we can think of a topic \( \phi_k \) as a generalised ingredient and a set of topics as ingredient bases: each recipe, but also each ingredient, becomes then a mixture of topics, i.e. of generalised ingredients [13]. A key feature of LDA is that the mixture of topics \( \pi(\phi_k) \) is itself a random variable with distribution \( p_k(\pi) \) over the topic space \( \mathbb{P}_k \) [2]. The likelihood of a string \( \{X_1, X_2, \ldots, X_n\} \) becomes then:

\[
P_{LDA} = \int_{\mathbb{P}_k} \left\{ \prod_{i=1}^{n} \sum_{\phi=1}^{k} \pi(\phi) T_\phi(X_i) \right\} p_k(\pi) d\pi.
\]

where \( T_\phi(X_i) \) is the probability of topic \( \phi_k \). The density function \( p_k(\cdot) \) is a Dirichlet distribution with parameters \( \eta_1 \ldots \eta_k \), that are usually clustered into two hyperparameters \( \alpha \) and \( \beta \) that represent document-topic (\( \alpha \)) and topic-word (\( \beta \)) distributions, respectively. Parameter \( \alpha \) roughly corresponds to topic leakage: small \( \alpha \) values will result in recipes that are mixtures of just few topics. On the other hand, \( \beta \) represents word leakage: high \( \beta \) values will induce topics as mixtures of most of the ingredients. For example, the \( T \)-topics representation with parameters \((\alpha, \beta = 0.01, T = 50)\) of recipe spaghetti met tomatensaus en worstjes (spaghetti in tomatosauce and sausages) is a mixture of 3 topics, with topic 37 being the most important:

Spaghetti in tomato sauce and sausages (smoked bacon, diced, Tuscan sausage, onion, garlic, sellery, carrot, diced tomatoes, spaghetti)

\[
\begin{align*}
&\text{Topic 1: } '\text{eggs, sugar, flour, whipped cream, flour, icing sugar, red pepper}' \\
&\text{Topic 2: } '\text{garlic, diced tomatoes, basil spinach, onion}' \\
&\text{Topic 3: } '\text{basil, mozzarella, cherrytomato, olive oil, extra virgin olive oil, rucola}'
\end{align*}
\]

for \( \alpha = 0.02 \), and:

\[
\begin{align*}
&\text{Topic 1: } '\text{carrot, onion, laurel, sellery, leek, beef broth}' \\
&\text{Topic 2: } '\text{olive oil, garlic, diced tomatoes, basil spinach, onion}' \\
&\text{Topic 3: } '\text{basil, mozzarella, cherrytomato, olive oil, extra virgin olive oil, rucola}'
\end{align*}
\]

for \( \alpha = 0.2 \), and finally for \( \alpha = 2 \) the same recipe is a mixture of topics \( \{5, 10, 14, 22, 34, 42, 45, 50\} \) all with weight 1.

Although generalised ingredients, being LDA topics, do not have necessarily a physical interpretation, qualitatively we found out that they correspond to ingredients sets typical of a particular dish type, e.g. a cake or a pasta. For example the top-5 topics for describing the AH dataset with \((\alpha = 10, \beta = 0.01, T = 5)\) are:

\[
\begin{align*}
&\text{TOPIC}_1: 0.15620: \text{eggs, butter, sugar, flour, sugar, rice, sugar, red pepper, flour, eggs} \\
&\text{TOPIC}_2: 0.21487: \text{oil, olive oil, onion, garlic, red pepper, diced tomatoes, tomatos, tomatos} \\
&\text{TOPIC}_3: 0.22453: \text{oil, garlic, diced tomatoes, basil spinach, onion} \\
&\text{TOPIC}_4: 0.23407: \text{oil, garlic, extra virgin olive oil, lemon, basil, rucola} \\
&\text{TOPIC}_5: 0.25993: \text{sunflower seed oil, garlic, olive oil, extra virgin olive oil, rucola}
\end{align*}
\]

Results

For evaluating whether substitution rules can be contextualised in the AH dataset we consider the task of finding substitutes in selected recipes, assessing whether we can rank candidate substitutes based on some metric defined on the topic space. Figure 1 schematically visualises this task.
Figure 1: The ranking task: ingredients from the CT rules, organised in substitution classes, are ranked by means of a statistical representation of the recipe, which the rule is being applied to. Recipe representations are a mixture of LDA topics or generalised ingredients.

Resulting rankings are self-judged and we choose LDA hyperparameters $\alpha = 0.2$, $\beta = 0.01$, $T = 50$. Consider the recipe:

Ciabattapizza = 'olive oil, mixed vegetables aubergine, ciabatta, sun dried tomato sauce, natural tuna, mozzarella'

and the substitution rules, obtained by mixing implicit, from the categorisation in CT, and explicit rules in the same source:

olive oil substitutes: all instances of the class Fats & Oils->Oils & Cooking Sprays
(tulsi oils are best used in cold dishes; heat destroys their delicate flavor.)
tuna substitutes swordfish OR sturgeon OR mako shark OR marlin OR halibut OR salmon OR mackerel
tomato sauce substitutes: tomato puree (this is unseasoned, while tomato sauce has salt, pepper, and other seasonings added) OR 3 parts tomato paste + 4 parts water (tomato paste has only salt added, while tomato sauce includes salt, pepper, and other seasonings.

The representation of this recipe in the topic space is

$$DP(1647,:) = 1x(1,37)+5x(1,48)$$

Topic 48 = '0.13086 olive oil, 0.10978 capers, 0.10929 garlic, 0.10243 black olives, 0.09312 anchovies, 0.04656 lemon'

Topic 37 = '0.13053 basil, 0.08750 mozzarella, 0.07507 cherry tomato, 0.06216 olive oil, 0.06073 e/v olive oil, 0.05260 rucola'

We can rank first candidate substitutes using the ingredients coefficients in each topic. For this example extra virgin olive oil ranks highest among the Oils & Cooking Sprays class, in line with our expectations since this is a cooking fat typical of the Italian cuisine and far less common elsewhere.

Composite ingredients, such as a tomato sauce might not be have one single substitute and we often need additional rules such as

Italian seasoning substitutes: equal parts basil, marjoram, oregano, rosemary, sage, and thyme.

for expanding 'other seasoning' here above. Basil, the top ingredient of topic 37 will then be ranked highest. Finer rankings can be generated by considering the closest topics according to the Kullback-Leibler [12] (KL) divergence,

Topic 46 = garlic, red onion, chili powder, parsley, tomato, oil
Topic 6 = olive oil, courgette, thyme, Parmesan cheese, rosemary, spinach
Topic 15 = e/v olive oil, white wine vinegar, red wine vinegar, mustard, red onion, shallot
Topic 4 = avocado, lemon juice, smoked chicken breast, leek, rucola, tomatoes
Topic 43 = leek, cream cheese, chive, mashed potato salmon fillet, margerine

These additional topics yield ranking values for more tuna substitutes (salmon fillet), seasonings (parsely, rosmary), among others.

Discussion and Future Work

In this work we qualitatively show that topic densities of document's mixtures and their neighbours in a KL divergence can be used for ranking candidate substitutes from expert generated rules. Notice that we can also start from the representation in the topic space and work out the substitution rules by organising ingredients based on their position in the ontology or in an explicit rule. The division of labour between statistical methods and food ontologies still holds. Intuitively, food ontologies are effective for discriminating on type on ingredients, which are difficult to distinguish in the topic space.

As the results are promising, we aim at extending this work in several ways. First, substitution rules should be evaluated extensively: on larger datasets and by means of user studies where recipes are prepared and tasted. Secondly, the basic food ontology derived from the CT thesaurus can be extended and possibly linked to other information sources. For example adding nutritional information can lead to more sophisticated substitution rules where both taste and nutrients are taken into account. Finally, explicit rules can be encoded in an ontology as well so that applications can query both sources transparently.
References