THEORIA ET HISTORIA SCIENTIARUM, VOL. VIII, N° 1
Ed. Nicolaus Copernicus University 2008

Tony Veale

Conceptual Refactoring for Creative Information Retrieval

Abstract

Information retrieval (IR) is an effective mechanism for text management that has received widespread adoption in the world at large. But it is not a particularly creative mechanism, in the sense of creating new conceptual structures or refactoring existing ones to pull in documents that describe, in novel and inventive ways, a user's information needs. Since language is a dynamic and highly creative medium of expression, the concepts that one seeks will therefore represent a moving target for IR systems. We argue that only by thinking creatively, and viewing concepts as fluid meaning structures capable of dynamic reorganization, can an IR system effectively retrieve documents that express themselves creatively.

1. Introduction

Most retrieval of textual information is literal in the sense that any retrieved document will literally match the keywords of the user's initial query. The query, whether a simple bag of conjoined keywords or a complex Boolean filter, essentially specifies the indices that should be examined to find matching documents. The set of matching documents is thus circumscribed by the keywords chosen by the user, making retrieval vulnerable to the *word mismatch problem* [2] if the authors of the most relevant documents have chosen to lexicalize their ideas in a different way. Of course, statistical and knowledge-based techniques (e.g., [3, 4, 5]) can be used to expand a query with highly correlated terms to permit the retrieval of additional relevant documents that do not literally contain any of the initial query terms. However, even these techniques still operate on the literal plane of meaning, by focusing on the conventional

meaning of the keywords used (e.g., by using their synonyms, hypernyms and hyponyms).

This literal mindset in information retrieval (IR) ignores the fact that language is a creatively dynamic medium, one that is always striving to find new ways to communicate the same old ideas, often with an additional connotation or a different spin [6 |. So while users of IR may be relentlessly literal in their choice of search terms, the authors of the documents they are hoping to retrieve will frequently be far more creative in their choice of words. To successfully retrieve these documents, it will be necessary for IR systems to demonstrate an equal level of creativity, to predict the innovative ways in which a relevant document might speak to the information needs of the user. These predictive techniques should be creative in the sense that they are capable of reorganizing an existing conceptual worldview (modeled using a taxonomy like WordNet [1], say) to look at a concept in new and interesting ways. We refer to this kind of reorganization as *refactoring*-, just as large numbers can be factorized in different ways, and software systems can be modularized around different abstraction hierarchies, so too can complex concepts be expressed using different combinations of simpler concepts. This refactoring is certainly creative in the sense that it requires an agile grasp of conceptual possibilities and would be considered ingenious if performed by a human, but it is also creative in the most obvious sense of the word, by exhibiting a capability to hypothesize and create new concepts that lead to insightful documents being retrieved.

Conceptual refactoring is the central challenge of what I like to call "creative information retrieval". Creative IR is all about augmenting traditional IR engines with the conceptual tools and representations needed to express themselves creatively, so that they can better predict the ways in which a user's search concept might be creatively communicated. The need for creative IR is perhaps most keenly felt by search engines that manage a relatively small corpus of documents with little or no redundancy, such as on-line product catalogs. In such situations, a user must anticipate the ways in which a product may be marketed by its creators and choose an appropriate query to retrieve for that product. For example, it is now intellectually fashionable to refer to certain *comic books* as *graphic novels.* In a similar fashion, *suntan lotion* is variously marketed as *suntan oil, suntan ointment, suntan gel, suntan cream, suntan milk* and even *suntan butter.* Some of these variations are predictable from literal knowledge of the lotion/ remedy domain, but others are clear uses of a food metaphor in which cream-as- lotion is perhaps the most entrenched instance. Product marketers strive for originality, so a statistical approach may not always learn such associations. However, in this paper we demonstrate how a creative system with basic metaphor capabilities can generate these variations from first principles, using a lexical knowledge-base like WordNet [1, 7].

2. Query Expansion

The search terms chosen by a user reflect the information needs of the user, but do not necessarily reflect the best set of indices with which to retrieve that information. Rather than use these terms as a query directly, intelligent search engines use them instead as merely a basis for constructing a query. This construction process, conventionally called query expansion, attempts to construct a rich query from the keywords offered by the user, in the hope that they will lead to greater document recall at equivalent levels of precision. Expansion of a user query can be performed using a variety of techniques, some of which are straightforward and mundane, but only some of which deserve to be labeled creative.

Statistical techniques can, in many cases, recognize domain correlations between terms, so that a query can be expanded with additional search terms that retrieve documents relevant to the same topic. For instance, corpus analysis reveals a strong co-occurrence probability for the word pairings *doctor* and *nurse*, *hospital* and *healthcare*, *Jaguar* and *sportscar*, etc. Relevance feedback techniques expand a query using terms extracted from the documents that a user has already marked as relevant to a query. The goal is to construct a query that returns more documents that are similar to those in which the user has expressed an interest, and dissimilar to those the user has deliberately avoided. Removing the need for user interaction, local context analysis (LCA) exploits the terms provided by the user as a first-cut query to retrieve an initial set of documents, the best of which are then statistically analysed to suggest additional search terms (both words and phrases) for an expanded query that will retrieve even more documents [3], Statistical techniques typically require no domain knowledge to operate, and most can adapt to word-usage trends automatically if they are continually trained as new documents enter the IR index.

In contrast, knowledge-based approaches use a domain model to recognize the concepts denoted by the user's search terms, which enables their associated definitions to be exploited. This domain model may be provided by a general purpose lexical ontology like WordNet, a broad-coverage structured lexicon for the English language. By pin-pointing particular entries in such an ontology, the expansion process can exploit existing knowledge about synonyms (e.g., WordNet considers one sense of *cream* to be a synonym for *ointment),* hypernyms and hyponyms (e.g., WordNet considers *holy oil* to be a hyponym of *ointment*, which suggests that *oil* might be a useful alternate for *ointment* in a query), as well as partonyms and holonyms (e.g., WordNet describes one sense of *gondola* as being part of an *airship).*

These techniques tend, by their nature, to derive a literal perspective on a query and its conceptual content. Few large-scale ontologies contain explicit knowledge about metaphors and the schematic structures that permit them to be comprehended. Additionally, since novelty is a driving force in creative thinking, statistical techniques will not find sufficient data from which to derive the associations needed to understand creative metaphors (though we acknowledge that there has been some success in statistically recognizing established conventional metaphors).

3. Concept Creation

One test of a creative system is its ability to hypothesize and create new concepts to suit a given situation. This test discriminates those systems that mimic creative behavior through the use of pre-coded rules and look-up tables from those that exhibit genuine innovation. For instance, a system might display an understanding of metaphoric language if it is given a sufficiently rich lexicon with appropriate cross-domain mappings, but one could not call such a system creative, as it limited to consider only those concepts that are defined by its knowledge-base. In contrast, a system that constructs such mappings on the fly, and which is thus not a priori limited, is creative in the sense of actually creating new knowledge.

Consider a simple case of knowledge-based query expansion for the user query *Jewish bread.* WordNet defines a range of hyponyms for {bread, breadstuff}, such as {muffin}, {wafer}, {biscuit} and {loaf}, that can all be used as expansion terms for *bread.* Additionally, WordNet specifies *Hebrew* as a synonym for *Jew,* so after some basic morphological analysis, we arrive at the following expansion:

Ql: (Jewish *or* Hebrew) *near* (bread *or* loaf *or* biscuit *or* muffin)

WordNet defines a variety of specific bread types as hyponyms of {bread, breadstuff}, and two of these actually contain the word *Jewish* in their gloss. The first, {Challah}, is unambiguous and makes an excellent stand-alone query expansion. The second, {Jewish-Rye}, literally contains the query term in its lexical form, but still contributes to the expansion in a non-trivial way as follows:

Q2: Challah *or* ((Jewish *or* Hebrew) *near* (bread *or* loaf *or* biscuit *or* rye *or ...))*

This use of hyponyms enriches the original query in a very effective way, but not all the relevant hyponyms in WordNet are recruited. For instance, the concept {matzo, matzoh, matzah, unleavened_bread} is also very relevant, but its gloss is tersely specified simply as *eaten at Passover.* However, WordNet defines the gloss for {Passover, Pesah, Pesach} as:

"a Jewish festival [...] celebrating the exodus of the Israelites from Egypt"

This may lead a creative system to construct a new concept, {Passover-bread}, to capture the implicit relationship between *Jewish* and *Matzo* and to unlock a range of productive expansion terms for our query as follows:

Q3: Matzo *or* Matzoh *or* Matzah *or* Challah
or ((Jewish *or* Hebrew or unleavened *or* Passover *or* Pesah *or ...) near*
(bread *or* loaf *or* biscuit *or* muffin *or* rye))

Further hypotheses might be inductively created on the basis of this concept [11], For instance, since Passover is defined as a kind of {religious-holiday} in WordNet, a creative system might hypothesize a more general concept, {Jewish- Holiday-Bread}, in the expectation that other Jewish holidays may be associated with a particular kind of bread even in the absence of explicit WordNet examples. For instance, it happens that Hanukkah is associated with a traditional honey bread made with figs, a fact unknown to both WordNet and the authors of this paper until revealed by the creative information retrieval system described in this paper.

4. Lexical Analogies

Analogical expressions use terms from one domain of discourse to allude to terms in another, systematically parallel domain of discourse [12], Analogy is thus useful when one knows of, or suspects, the existence of a given concept but does not know how it is lexicalized. For instance, a user may know that Islam is based on a particular sacred text but not know what it is called or how it is spelled. In this case, an analogy like *the bible of Islam* can be used to allude to it indirectly. In a knowledge-based IR system, a lexical ontology like WordNet can be used to resolve this analogy prior to query expansion. The query that is then generated is simply:

Q4: Koran *or* Quran *or* (Islam *near* bible)

Lexical analogies do not need to generate a one-to-one mapping of concepts, but can involve an indeterminacy that may usefully increase recall in an information retrieval setting. For instance, there is no particular book that one can definitively term the *Hindu bible.* However, WordNet defines {bible, scripture} as a hypernym of {sacred-text, sacred-writing}, and there are several other hyponyms of this parent whose gloss mentions *Hindu*, such as {Brahmana}, {Veda}, {Samhita} and {Mahabharata}. This leads to the following query expansion:

Q5: Brahmana *or* Veda *or* Samhita *or* Mahabharata *or* "Hindu bible"

Such analogies are not difficult to generate using a lexical knowledge-base like WordNet, so this creative ability can be readily harnessed in an IR setting. Before explaining how, we must first consider some basic terminology. We define the *pivot* of a concept as the lowest hypernym that can be lexicalized as a single atomic term. The pivot of {Mars} is therefore {deity, god}, and not its immediate hypernym {Roman-deity}, which can only be lexicalized as a compound term. The intuition here is that compound terms are likely to represent a domain specialization of a concept, but good analogies should search outside this home domain to find a counterpart in a different part of the ontology. Additionally, we define the *discrimination set* of a concept as the set of words in its gloss that have been used in other WordNet compound terms as modifiers. For example, the gloss for {Mars} is shown below:

{Mars} "(Roman mythology) god of war and agriculture"

The discrimination set of {Mars} is *{war, agriculture}*, since *war* is used in WordNet to differentiate {crime, law-breaking} into {war-crime} and *agriculture* is used to differentiate {department, section} into {agriculture-section}. These lexical precedents suggest that *war* and *agriculture* might also be useful in discriminating the {deity, god} category since they are used to define {Mars} and may help us find a similar deity. We do not place *Roman* in this set because it is explicitly tied to the home domain of {Mars} via {Roman-deity} and thus has no analogical potential.

We use a combination of the pivot and the discrimination set to gather a set of candidate source concepts for the target. The candidate set is defined as the collection of all concepts that are hyponyms of the pivot (a similarity constraint), and which reside at the same depth of the ontology as the target (a specificity constraint), and whose gloss contains at least one of the terms in the discrimination set of the target (a relevance constraint). The following concepts are therefore considered plausible candidate sources for the analogical target concept {Mars}:

{Durga}	"{Hindu} goddess of war"
{Ares}	"(Greek mythology) god of war"
{Morrigan}	"(Irish) war goddess"
{Skanda}	"[Hindu] god of war"
(Ishtar, Mylitta}	"(Babylonian and Assyrian) goddess of love [...] and war"
{Tiu}	"[Anglo-Saxon) god of war and sky"
{Tyr, Tyrr}	"god of war and strife and son of Odin"
{Hachiman}	"[Japanese] Shinto god of war"
{Nabu, Nebo}	"(Babylonian) god of wisdom and agriculture ..."
{Brigit}	"(Irish) goddess of fire and fertility and agriculture ..."
{Dagan}	"(Mesopotamia) god of agriculture and earth"
{Dagon}	"(Phoenician and Philistine) god of agriculture and the earth"

From these candidates an analogical query expansion can be created as follows:

Q6: Roman *and* (Durga *or* Ares *or* Morrigan *or* Skanda *or* Ishtar *or* Tiu *or* Tyr *or* Hachiman *or* Nabu *or* Nebo *or* Brigit *or* Dagan *or* Dagon)

Although fishing for poetic references, this query does succeed in finding documents that allude to {Mars} without actually mentioning *Mars.* For example, book VII of the Sibylline Oracles contains the following quotation:

But to them afterwards
Shall Roman Ares flash from many a spear;

Though this example is whimsical, lexical analogies can run the gamut from the poetic to the technical. For a consideration of the IR potential of analogy, see [12, 13],

5. Lexical Metaphors and Polysemy

Metaphor is a highly-generative conceptual phenomenon that can be used to create a wide range of linguistic expressions that refer to the same concept [6], If we wish to retrieve documents that allude to a search concept figuratively rather than literally, it will be necessary to use an understanding of the metaphor process to expand the search query with plausible figurative lexicalizations of this concept. Consider the user query *Microsoft Monopoly.* There are three senses of *Monopoly* in WordNet:

{monopoly}: a hyponym of {dominance, control, ascendance)
"exclusive control or possession of something"

{monopoly}: a hyponym of (market, market-place}
"(Economics) A market in which there are many buyers but only one seller'' {monopoly}:

a hyponym of {board-game}

"(Trademark) A board-game in which players attempt to gain a monopoly..."

Clearly the first two senses are semantically related. Indeed, the second sense can be seen as a semantic bleaching of the first, in which the notion of exclusivity is abstracted out of the realm of economics and allowed to apply to any domain at all. We can exploit this polysemy to reformulate the idea of a monopoly as *a state of dominance in a market-place,* which in turn yields the expansion of Q7 below:

Q7: Microsoft *and* (monopoly *or* (market *near* (control *or* dominance))

This reformulation offers a more coherently unified view of the concept Monopoly than any single sense stored in WordNet. This should not be too surprising,

since polysemy is often an artificial side-effect of the violence a sense- differentiating lexicon like WordNet must do to a concept in order to fit it into a branching taxonomic structure. Distinct word senses in WordNet thus capture just a sliver of the relational structure of the psychological concept, and it requires some creative thinking to reconstitute the conceptual whole from its various sense fragments. The reformulation in Q7 is a purely literal one, but it is also a more explanatory one from the perspective of IR, as expansion Q7 will now tend to retrieve documents that actually explain the hypothesis that Microsoft has a monopoly, by referring to dominance in particular markets, and should give more weight to those documents than to those that simply echo the hypothesis without additional insight.

By opening up a concept into a conjunction of other concepts, polysemy exposes the inner components of conceptual meaning to greater scrutiny, so that if applied recursively, it allows us to reformulate a metaphoric definition of the concept. For instance, the concept {dominance, control, ascendance} has other hyponyms besides {monopoly}, such as (mastery, supremacy}, {predominance, predomination}, (rule, domination} and, most interestingly, {tyranny, despotism}. This enables the following metaphoric expansion:

> Q8: Microsoft *and* (monopoly *or* (market *near* (dominance *or* rule *or* mastery *or* tyranny *or* despotism *or* ascendance *or* supremacy))

In turn, *despotism* has two related senses in WordNet, the sense in Q8 which is a hyponym of (dominance, control} and another, more intense sense from the political domain: {despotism, totalitarianism, Stalinism, authoritarianism, dictatorship}. This sense leads to a highly-charged metaphoric expansion that should retrieve documents conveying the more extremist perspectives on Microsoft's market position.

Note how the concept {market, market-place} is not figuratively extended in Q8. The key to using polysemy for metaphoric query expansion is that only one component of a polysemous sense pair should be extended, while the other is left unchanged to serve as a domain anchor for the expansion, ensuring that the query remains on topic. Thus in Q7 and Q8, colorful terms like *tyranny* are constrained to occur in the proximity of the anchor term *market*, ensuring that the query never strays from its original focus. Conversely, if {market, marketplace} is extended, to provide the term *shelf* say, {dominance, control, ascendance} should not.

These expansion strategies assume that instances of polysemy can be recognized in WordNet and differentiated from instances of homonymy, another form of lexical ambiguity in which the senses of a word are not psychologically related. However, the distinction between each kind of ambiguity is not explicitly marked in WordNet.

5.1. Detecting Polysemy in WordNet

Nonetheless, polysemous relationships can be recognized using a variety of automatic approaches. In the top down approach, *cousin relations* [7, 8] are manually established between concepts in the upper-ontology to explain the systematicity of polysemy at lower levels. For instance, once a connection between {animal} and {food} is established, it can be instantiated by words with both an animal and a food sense, such as *chicken* and *lamb.* This approach is limited by the number of high-level connections that are manually added, and by the need to list often copious exceptions to the pattern (e.g., *mate* the animal partner, and *mate* the berry drink, are merely homonyms; the latter is not derived from the former). Conversely, in the bottom-up approach, systematic patterns are first recognized in the lower ontology and then generalized to establish higher-level connections [9, 10]. For instance, several words have senses that denote both a kind of music and a kind of dance (e.g., *waltz, tango, conga),* which suggests a polysemous link between {music} and {dance}.

Both of these approaches treat polysemy as a systematic phenomenon best described at the level of word families. However, while such a treatment reveals interesting macro-tendencies in the lexicon, it does little to dispel the possibility that homonymy might still operate on the micro-level of individual words (as demonstrated by the size of the exception list needed for the first approach). We prefer instead to use an evidential case-by-case approach to detecting polysemy, connecting a pair of senses only when explicit local taxonomic evidence can be found to motivate a connection. This evidence can take many forms, so a patchwork of heuristic detectors is required. We describe here the three most interesting of these heuristics.

Explicit Ontological Bridging: a sense pair co,> for a word co can be linked if co $_t$ has a hypernym that can be lexicalized as M-H and co, has a hypernym that can be lexicalized as M, the rationale being that co, is the M of co j and co$_x$ is the H of co,. E.g., the word *olive* has a sense with a hypernym {fruit-tree}, and another with the hypernym {fruit}, therefore M = fruit and H = tree. *(Coverage:* 72%, *Accuracy: 94%).*

Hierarchical Reinforcement: if <oq, a,> and <P,, p,> are sense pairs for two words a and P where oq is a hypernym of P(and cq is a hypernym of P,, then <oq, oc,> reinforces the belief that <P,, P,> is polysemous, and vice versa. For example, the word *herb* denotes both a plant and a foodstuff in WordNet, and each of these senses has a hyponym that can be lexicalized as *sage. (Coverage:* 7%, *Accuracy: 12%).*

Cross-Reference: if $<\omega_1, \omega_2>$ is a sense pair for a word w and the WordNet gloss for ω_2 explicitly mentions a hypernym of ω_1, then ω_2 can be seen as a conceptual extension of ω_1. For instance, the railway-compartment sense of *diner* mentions *restaurant* in its gloss, while another sense actually specihes {restaurant} as a hypernym. This suggests that the railway sense is an extension of the restaurant sense that uses the later as a ground for its definition. *(Coverage: 62%, Accuracy: 85%).*

The coverage of each heuristic is estimated relative to that achieved by the *cousins* collection of 105 regular polysemy noun-sense groupings that are hand- coded in WordNet [7, 8], Over-generation is estimated relative to the overlap with the cousins exception list [7], which permits us to also estimate the accuracy of each heuristic.

6, Concept Combination

User keyword-combinations run the gamut from existing WordNet compounds like *drug addict* to novel extensions of these compounds. The former are trivial to handle, and even the latter are straightforward if they can be shown to taxonomically extend an existing compound. In the case of *Prozac addiction,* the WordNet entry {drug-addiction} can be used as a guide to interpretation, since WordNet already defines {Prozac} as a kind of {drug}. Knowing how a novel compound fits into the WordNet taxonomy is extremely valuable for query expansion purposes. For instance, WordNet defines other hyponyms of {drug-addiction}, such as {heroin-addiction}, {cocaine-addiction} and {alcohol- addiction}, that also follow the same instantiation pattern since *heroin, cocaine* and *alcohol* all denote a hyponym of {drug}. These hyponyms point to other concepts that, by virtue of being lexicalized as modifiers of *addiction,* serve as prototypical addictive drugs. This allows us to generate a query composed of similes in which addiction is implicit:

Q9: Prozac *near* like *near* (cocaine *or* heroin *or* opium *or* nicotine *or* alcohol)

These prototypes also point to other concepts that are strongly suggestive of addiction. For instance, in addition to the {drug-addiction}/}heroin-addiction} instantiation pattern, WordNet also instantiates {drug-addict, junky, junkie} with {heroin-addict} and {drug-abuse, habit} with {alcohol-abuse}. This leads us to consider other topics that are strongly related to the search query such as *Prozac addict* and *Prozac abuse.* Additionally, these compounds suggest useful synonyms in which the notion of addiction is implicit, such as *habit, junky* and *junkie,* but which do not arise from a consideration of the concept {addiction, dependency} in isolation.

Expansions such as these are a product of metonymy, a conceptual mechanism whereby a concept is used as a referential proxy for another, strongly associated concept. The textual glosses in WordNet are a fruitful basis for exploiting metonymy between concepts. For instance, from {drug-addict} one can infer a metonymic link to {withdrawal-symptom}, since the gloss for the latter contains the phrase *drug addict.* This metonymy suggests that *Prozac withdrawal symptom* is a relevant topic to include in the overall expansion, as follows:

Q10: Prozac *near* ((like *near* (cocaine *or* heroin *or* opium *or* nicotine *or* alcohol)) *or* (addict *or* addiction *or* dependency *or* habit *or* abuse *or* "withdrawal symptom"))

It is generally safe to expand compound terms in this way, because these are far less ambiguous in WordNet than atomic terms, with most having just a single sense.

7. Multi-Term Conceptual Refactoring

Concept combination, analogy and metaphor are all forms of conceptual refactoring, since each seeks to rearrange the underlying components of meaning in a concept. But a more explicit form of refactoring, closest in form to arithmetic refactoring and even software refactoring, can be seen in the treatment of multi- term phrases. For in trying to paraphrase these word combinations in creative ways, it is often necessary to shift elements of meaning from one term to another without affecting the meaning of the whole.

Consider the user query *Italian recipes.* Using term associations common to both query words, derived from either a semantic network or a statistical language model, a system can easily generate a query like the following:

Qll:(Italian *or* Mozzarella *or* pizza *or* antipasto) *near* recipes

The concept *Italian* is implicit in the concept *Mozzarella, Pizza* and *Antipasto,* and so we have not so much refactored the query as specialized it. But now consider how different associations can be chained together to yield a more creative expansion. WordNet 1.6 provides a strangely food-neutral definition of {recipe, formula} as *directions for making something,* and the only concept in which *recipe* appears as a gloss term is {cookbook, cookery-book}: *a book of recipes and cooking directions.* This suggests that *cookery-book* is actually a very effective metonym of *recipes,* so we can reformulate our query as:

Q12: (Italian *or* Mozzarella *or...) near* ("cookery book" *or* cookbook *or* recipes)

Some linguistic word-play allows us to go from *Italian cookery-book* to *Italian- cookery book* without loss of information. However, the latter proves to be a much more fruitful organization of the query. WordNet defines {cookery, cooking, cuisine, culinary-art} as *the practice or manner of preparing food or the food so prepared,* and only one of these gloss words, *food,* denotes a concept with hyponyms whose glosses contain *Italian.* Among those hyponyms of {food, nutrient} whose glosses mention *Italian* are {pizza}, {antipasto}, {Mozzarella}, {Ricotta} and {Frittata}. This reorganization thus suggests the following expansion:

Q13: ((Italian *and* (cookery *or* food)) *or* pizza *or* antipasto *or* Mozzarella or...) *near* (book *or* cookbook *or* recipes))

Now, in searching the conceptual vicinity of {cookery-book, cookbook} for potential metaphors, we recognize an interesting example of WordNet polysemy in the word *bible,* which can denote either an authoritative handbook or a sacred text. This polysemy suggests that at least one sense of *bible* is metaphoric, which in turn suggests that *bible* may be a good metaphor for a cookery book, especially since both {cookery-book} and {bible} share a very specific common hypernym in {reference-book}. This leads us to produce the following figurative expansion:

Q14: ((Italian *and* (cookery *or* food)) *or* pizza *or* antipasto *or* Mozzarella or...) *near* (book *or* cookbook *or* recipes *or* bible))

This final query is capable of matching creative allusions in text, such as *The Antipasto Bible,* that are strongly suggestive of the search concept *Italian recipes.* This suggestion is an emergent one of course, arising out of a delicate interaction of connotations from the words *Antipasto* and *Bible.*

The query fragment *Antipasto Bible* amply demonstrates what it means to refactor a conceptual structure, showing that meaning can be removed from one content component only if it is added to another to preserve the meaning of the whole. This need for meaning preservation explains why it would it be incoherent to simply leap from *recipes* to *Bible* without first discharging the query's responsibility to somehow capture the implicit *food* theme of the query. So in Q14 the term *Antipasto* contributes the meaning elements *Italian* and *food* while the term *Bible* contributes *book,* with both together contributing the requirec element *recipes* only as an emergent by-product, via *cookery-book,* of the combination *food + book.* The loss of meaning inherent in the refactoring of *recipes* into *book* is thus compensated by the increase in specificity arising from the refactoring of *Italian* into *Antipasto.*

8. Concluding Remarks

The mapping between language and conceptualization is a fluid one that provides considerable freedom to creative individuals to express themselves and their thoughts in inventively non-literal ways. The challenge of creative IR then is to be able to explore and stretch the boundaries of linguistic expression in a similar manner, while simultaneously keeping an expanded user-query firmly on-topic. While this requires a practical perspective on creative language processing, the ideas presented in this paper are nonetheless founded on the belief that creativity is fundamentally a conceptual phenomenon that demands explicit conceptual representation, rather than implicit modeling via the statistical properties of language. Of course, it is certainly true that statistical models of conceptual association are possible and can improve IR performance (e.g., see [2], [3]). These techniques also have the considerable advantage of being dynamically attuned to the changing trends of language use. It is also true that, in contrast to statistical models, knowledge-based systems can involve complex tangles of axioms and rules that may need frequent maintenance and sanity-checking by ontological engineers.

Nonetheless, when it comes to chaining together successive associations to generate a truly creative leap of the imagination, we believe that only a knowledge- based approach can sufficiently ensure the logical coherence of the resulting expansion. Our discussion of conceptual refactoring illustrates the book-keeping aspects of query expansion, whereby meaning elements can be moved from one content term to other, thus demonstrating that an explicit logical model is needed to ensure a balanced result. Such book-keeping cannot be achieved by statistical techniques alone. Furthermore, only a knowledge-based approach is capable of representing and reasoning about metaphor as a deep conceptual phenomenon, rather than as an epiphenomenon of word distribution. Statistical techniques will undoubtedly have a non-trivial role to play, but for the reasons we outline in this paper, we believe the challenge of creative information retrieval is one that will be achieved only within the context of a knowledge-based framework.

References

1. Miller, G. A. WordNet: A Lexical Database for English. Communications of the ACM, Vol. 38 No. 11 (1995)
2. Furnas, G. W., Landauer, T. K., Gomez, L. M., Dumais, S. T. The vocabulary problem in human-system communication. Communications of the ACM, vol. 30, number 11. (1987)

3. Xu, J., Croft, W. B. Improving the Effectiveness of Information Retrieval using Local Context Analysis. ACM Transactions on Information Systems, vol. 18, number 1. (2000)
4. Qiu, Y., Frei, H. P. Concept-based Query Expansion. In the proceedings of the ACM SIGIR Int. Conference on Research and Development in Information Retrieval. (1993)
5. Navigli, R., Velardi, P. An Analysis of Ontology-Based Query Expansion Strategies. In the proceedings of the Int. Workshop on Adaptive Text Extraction and Mining at the 14lh European Conference on Machine Learning. Dubrovnik, Croatia. (2003)
6. Lakoff, G, Johnson, M. Metaphors We Live By. Uni. of Chicago Press: Chicago (1980)
7. WordNet documentation, www.princeton.edu/~wn/ (2003)
8. Peters, W., Peters, I., Vossen, P. Automatic sense clustering in Euro WordNet. In the proceedings of the 1st international conference on Language Resources and Evaluation. Spain (1998)
9. Peters, I., Peters, P. Extracting Regular Polysemy Patterns in WordNet. Technical Report, University of Sheffield, UK. (2000)
10. Peters, W., Peters, I. Lexicalized Systematic Polysemy in WordNet. In the proceedings of the 2nd international conference on Language Resources and Evaluation. Athens. (2000)
11. Colton, S. Creative Logic Programming. In the proceedings of the 3rd Workshop on Creative Systems, held as part of the 18th International Joint Conference on Artificial Intelligence, IJCAI'03, Acapulco, Mexico (2003)
12. Tony Veale. The Analogical Thesaurus: An Emerging Application at the Juncture of Lexical Metaphor and Information Retrieval. In the proceedings of IAAI 2003, the 2003 International Conference on Innovative Applications of Artificial Intelligence. Acapulco, Mexico (2003)
13. Tony Veale. Dynamic Type Creation in Metaphor Interpretation and Analogical Reasoning: A Case-Study with WordNet. In the proceedings of ICCS2003, the 2003 International Conference on Conceptual Structures, Dresden, Germany (2003)